Low-Rank Covariance Function Estimation for Multidimensional Functional Data

Jiayi Wang[1], Raymond K. W. Wong*[1], and Xiaoke Zhang†[2]

[1]*Department of Statistics, Texas A&M University*
[2]*Department of Statistics, George Washington University*

September 1, 2020

Abstract

Multidimensional function data arise from many fields nowadays. The covariance function plays an important role in the analysis of such increasingly common data. In this paper, we propose a novel nonparametric covariance function estimation approach under the framework of reproducing kernel Hilbert spaces (RKHS) that can handle both sparse and dense functional data. We extend multilinear rank structures for (finite-dimensional) tensors to functions, which allow for flexible modeling of both covariance operators and marginal structures. The proposed framework can guarantee that the resulting estimator is automatically semi-positive definite, and can incorporate various spectral regularizations. The trace-norm regularization in particular can promote low ranks for both covariance operator and marginal structures. Despite the lack of a closed form, under mild assumptions, the proposed estimator can achieve unified theoretical results that hold for any relative magnitudes between the sample size and the number of observations per sample field, and the rate of convergence reveals the phase-transition phenomenon from sparse to dense functional data. Based on a new representer theorem, an ADMM algorithm is developed for the trace-norm regularization. The appealing numerical performance of the proposed estimator is demonstrated by a simulation study and the analysis of a dataset from the Argo project.

Keywords: Functional data analysis; multilinear ranks; tensor product space; unified theory

1 Introduction

In recent decades, functional data analysis (FDA) has become a popular branch of statistical research. General introductions to FDA can be found in a few monographs (e.g., Ramsay and Silverman, 2005; Ferraty and Vieu, 2006; Horváth and Kokoszka, 2012; Hsing and Eubank, 2015; Kokoszka and Reimherr, 2017). While traditional FDA deals with a sample of time-varying trajectories, many new forms of functional data have emerged due to improved capabilities of data recording and storage, as well as advances in scientific computing. One particular new form of

*The research of Raymond K. W. Wong is partially supported by National Science Foundation grants DMS-1806063, DMS-1711952 and CCF-1934904.

†The research of Xiaoke Zhang is partially supported by National Science Foundation grant DMS-1832046.

functional data is *multidimensional functional data*, which becomes increasingly common in various fields such as climate science, neuroscience and chemometrics. Multidimensional functional data are generated from random fields, i.e., random functions of several *input* variables. One example is multi-subject magnetic resonance imaging (MRI) scans, such as those collected by the Alzheimer's Disease Neuroimaging Initiative. A human brain is virtually divided into three-dimensional boxes called "voxels" and brain signals obtained from these voxels form a three-dimensional functional sample indexed by spatial locations of the voxels. Despite the growing popularity of multidimensional functional data, statistical methods for such data are limited apart from very few existing works (e.g., Huang et al., 2009; Allen, 2013; Zhang et al., 2013; Zhou and Pan, 2014; Wang and Huang, 2017).

In FDA covariance function estimation plays an important role. Many methods have been proposed for unidimensional functional data (e.g., Rice and Silverman, 1991; James et al., 2000; Yao et al., 2005; Paul and Peng, 2009; Li and Hsing, 2010; Goldsmith et al., 2011; Xiao et al., 2013), and a few were particularly developed for two-dimensional functional data (e.g., Zhou and Pan, 2014; Wang and Huang, 2017). In general when the input domain is of dimension p, one needs to estimate a $2p$-dimensional covariance function. Since covariance function estimation in FDA is typically nonparametric, the curse of dimensionality emerges soon when p is moderate or large.

For general p, most work are restricted to regular and fixed designs (e.g., Zipunnikov et al., 2011; Allen, 2013), where all random fields are observed over a regular grid like MRI scans. Such sampling plan leads to a tensor dataset, so one may apply tensor/matrix decompositions to estimate the covariance function. When random fields are observed at irregular locations, the dataset is no longer a completely observed tensor so tensor/matrix decompositions are not directly applicable. If observations are densely collected for each random field, a two-step approach is a natural solution, which involves pre-smoothing every random field followed by ensor/matrix decompositions at a fine discretized grid. However, this solution is infeasible for sparse data where there are a limited number of observations per random field. One example is the data collected by the international Argo project (`http://www.argo.net`). See Section 7 for more details. In such sparse data setting, one may apply the local smoothing method of Chen and Jiang (2017), but it suffers from the curse of dimensionality when the dimension p is moderate due to a $2p$-dimensional nonparametric regression.

We notice that there is a related class of literature on longitudinal functional data (e.g., Chen and Müller, 2012; Park and Staicu, 2015; Chen et al., 2017), a special type of multidimensional functional data where a function is repeatedly measured over longitudinal times. Typically multi-step methods are needed to model the functional and longitudinal dimensions either separately (one dimension at a time) or sequentially (one dimension given the other), as opposed to the joint estimation procedure proposed in this paper. We also notice a recent work on longitudinal functional data under the Bayesian framework (Shamshoian et al., 2019).

The contribution of this paper is three-fold. First, we propose a new and flexible nonparametric method for low-rank covariance function estimation for multidimensional functional data, via the introduction of (infinite-dimensional) unfolding operators (See Section 3). This method can handle both sparse and dense functional data, and can achieve joint structural reductions in all dimensions, in addition to rank reduction of the covariance operator. The proposed estimator is guaranteed to be semi-positive definite. As a one-step procedure, our method reduces the theoretical complexities compared to multi-steps estimators which often involve a functional principal component analysis followed by a truncation and reconstruction step (e.g., Hall and Vial, 2006; Poskitt and

Sengarapillai, 2013).

Second, we generalize the representer theorem for unidimensional functional data by Wong and Zhang (2019) to the multidimensional case with more complex spectral regularizations. The new representer theorem makes the estimation procedure practically computable by generating a finite-dimensional parametrization to the solution of the underlying infinite-dimensional optimization.

Finally, a unified asymptotic theory is developed for the proposed estimator. It automatically incorporates the settings of dense and sparse functional data, and reveals a phase transition in the rate of convergence. Different from existing theoretical work heavily based on closed-form representations of estimators, (Li and Hsing, 2010; Cai and Yuan, 2010; Zhang and Wang, 2016; Liebl, 2019), this paper provides the first unified theory for penalized global M-estimators of covariance functions which does not require a closed-form solution. Furthermore, a near-optimal (i.e., optimal up to a logarithmic order) one-dimensional nonparametric rate of convergence is attainable for the $2p$-dimensional covariance function estimator for Sobolev-Hilbert spaces.

The rest of the paper is organized as follows. Section 2 provides some background on reproducing kernel Hilbert space (RKHS) frameworks for functional data. Section 3 introduces Tucker decomposition for finite-dimensional tensors and our proposed generalization to tensor product RKHS operators, which is the foundation for our estimation procedure. The proposed estimation method is given in Section 4, together with an computational algorithm. The unified theoretical results are presented in Section 5. The numerical performance of the proposed method is evaluated by a simulation study in Section 6 and a real data application in Section 7.

2 RKHS Framework for Functional Data

In recent years there is a surge of RKHS methods in FDA (e.g., Yuan and Cai, 2010; Zhu et al., 2014; Li and Song, 2017; Reimherr et al., 2018; Sun et al., 2018; Wong et al., 2019). However, covariance function estimation, a seemingly well-studied problem, does not receive the same amount of attention in the development of RKHS methods, even for unidimensional functional data. Interestingly, we find that the RKHS modeling provides a versatile framework for both unidimensional and multidimensional functional data.

Let X be a random field defined on an index set $\mathcal{T} \subset \mathbb{R}^p$, with a mean function $\mu_0(\cdot) = \mathbb{E}\{X(\cdot)\}$ and a covariance function $\gamma_0(*, \cdot) = \mathrm{Cov}(X(*), X(\cdot))$, and let $\{X_i : i = 1, \ldots, n\}$ be n independently and identically distributed (i.i.d.) copies of X. Typically, a functional dataset is represented by $\{(\boldsymbol{T}_{ij}, Y_{ij}) : j = 1, \ldots m_i; i = 1, \ldots, n\}$, where

$$Y_{ij} = X_i(\boldsymbol{T}_{ij}) + \epsilon_{ij} \in \mathbb{R} \tag{1}$$

is the noisy measurement of the i-th random field X_i taken at the corresponding index $\boldsymbol{T}_{ij} \in \mathcal{T}$, m_i is the number of measurements observed from the i-th random field, and $\{\epsilon_{ij} : i = 1, \ldots, n; j = 1, \ldots m_i\}$ are independent errors with mean zero and finite variance. For simplicity and without loss of generality, we assume $m_i = m$ for all i.

As in many nonparametric regression setups such as penalized regression splines (e.g., Pearce and Wand, 2006) and smoothing splines (e.g., Wahba, 1990; Gu, 2013), the sample field of X, i.e., the realized X (as opposed to the sample path of a unidimensional random function), is assumed to reside in an RKHS $\mathcal{H}$ of functions defined on $\mathcal{T}$ with a continuous and square integrable reproducing kernel K. Let $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and $\| \cdot \|_{\mathcal{H}}$ denote the inner product and norm of $\mathcal{H}$ respectively. With the technical condition $\mathbb{E}\|X\|_{\mathcal{H}}^2 < \infty$, the covariance function γ_0 resides in the tensor product RKHS $\mathcal{H} \otimes \mathcal{H}$. It can be shown that $\mathcal{H} \otimes \mathcal{H}$ is an RKHS, equipped with the reproducing kernel $K \otimes K$

defined as $(K \otimes K)((\boldsymbol{s}_1, \boldsymbol{t}_1), (\boldsymbol{s}_2, \boldsymbol{t}_2)) = K(\boldsymbol{s}_1, \boldsymbol{s}_2)K(\boldsymbol{t}_1, \boldsymbol{t}_2)$, for any $\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{t}_1, \boldsymbol{t}_2 \in \mathcal{T}$. This result has been exploited by Cai and Yuan (2010) and Wong and Zhang (2019) for covariance estimation in the unidimensional setting.

For any function $f \in \mathcal{H} \otimes \mathcal{H}$, there exists an operator mapping $\mathcal{H}$ to $\mathcal{H}$ defined by $g \in \mathcal{H} \mapsto \langle f(*, \cdot), g(\cdot)\rangle_{\mathcal{H}} \in \mathcal{H}$. When f is a covariance function, we call the induced operator a $\mathcal{H}$-covariance operator, or simply a covariance operator as below. To avoid clutter, the induced operator will share the same notation with the generating function. Similar to L^2-covariance operators, the definition of an induced operator is obtained by replacing the L^2 inner product by the RKHS inner product. The benefits of considering this operator have been discussed in Wong and Zhang (2019). We also note that a singular value decomposition (e.g., Hsing and Eubank, 2015) of the induced operator exists whenever the corresponding function f belongs to the tensor product RKHS $\mathcal{H} \otimes \mathcal{H}$. The idea of induced operator can be similarly extended to general tensor product space $\mathcal{F}_1 \otimes \mathcal{F}_2$ where $\mathcal{F}_1$ and $\mathcal{F}_2$ are two generic RKHSs of functions.

For any $\gamma \in \mathcal{H} \otimes \mathcal{H}$, let $\gamma^\top$ be the transpose of γ, i.e., $\gamma^\top(\boldsymbol{s}, \boldsymbol{t}) = \gamma(\boldsymbol{t}, \boldsymbol{s})$, $\boldsymbol{s}, \boldsymbol{t} \in \mathcal{T}$. Define $\mathcal{M} = \{\gamma \in \mathcal{H} \otimes \mathcal{H} : \gamma \equiv \gamma^\top\}$. To guarantee symmetry and positive semi-definiteness of the estimators, Wong and Zhang (2019) adopted $\mathcal{M}^+ = \{\gamma \in \mathcal{M} : \langle \gamma f, f\rangle_{\mathcal{H}} \geq 0, \forall f \in \mathcal{H}\}$ as the hypothesis class of γ_0 and considered the following regularized estimator:

$$\arg\min_{\gamma \in \mathcal{M}^+} \{\ell(\gamma) + \tau\Psi(\gamma)\}, \tag{2}$$

where ℓ is a convex and smooth loss function characterizing the fidelity to the data, $\Psi(\gamma)$ is a spectral penalty function (see Definition 5 below), and τ is a tuning parameter. Due to the constraints specified in $\mathcal{M}^+$, the resulting covariance estimator is always positive semi-definite. In particular, if the spectral penalty function $\Psi(\gamma)$ imposes the trace-norm regularization, an ℓ_1-type shrinkage penalty on the respective singular values, the estimator is usually of low rank. Cai and Yuan (2010) adopted a similar objective function as in (2) but with the hypothesis class $\mathcal{H} \otimes \mathcal{H}$ and an ℓ_2-type penalty $\Psi(\gamma) = \|\gamma\|^2_{\mathcal{H}\otimes\mathcal{H}}$, so the resulting estimator may neither be positive semi-definite nor low-rank.

Although Cai and Yuan (2010) and Wong and Zhang (2019) focused on unidimensional functional data, their frameworks can be directly extended to the multidimensional setting. Explicitly, similar to (2), as long as a proper $\mathcal{H}$ for the random fields with dimension $p > 1$ is selected, an efficient "one-step" covariance function estimation with the hypothesis class $\mathcal{M}^+$ can be obtained immediately, which results in a positive semi-definite and possibly low-rank estimator. Since an RKHS is identified by its reproducing kernel, we simply need to pick a multivariate reproducing kernel K for multidimensional functional data. However, even when the low-rank approximation/estimation is adopted (e.g., by trace-norm regularization), we still need to estimate several p-dimensional eigenfunctions nonparametrically. This curse of dimensionality calls for a more efficient modeling. Below, we explore this through the lens of tensor decomposition in finite-dimensional vector spaces and its extension to infinite-dimensional function spaces.

3 Low-Rank Modeling via Functional Unfolding

In this section we will extend the well-known Tucker decomposition for finite-dimensional tensors to functional data, then introduce the concept of functional unfolding for low-rank modeling, and finally apply functional unfolding to covariance function estimation.

3.1 Tucker decomposition for finite-dimensional tensors

First, we give a brief introduction to the popular Tucker decomposition (Tucker, 1966) for *finite-dimensional* tensors. Let $\mathcal{G} = \bigotimes_{k=1}^{d} \mathcal{G}_k$ denote a generic tensor product space with finite-dimensional $\mathcal{G}_k, k = 1, \ldots, d$. If the dimension of $\mathcal{G}_k$ is q_k, $k = 1, \ldots, d$, then each element in $\mathcal{G} = \bigotimes_{k=1}^{d} \mathcal{G}_k$ can be identified by an array in $\mathbb{R}^{\prod_{j=k}^{d} q_k}$, which contains the coefficients through an orthonormal basis. By Tucker decomposition, any array in $\mathbb{R}^{\prod_{k=1}^{d} q_k}$ can be represented in terms of n-mode products as follows.

Definition 1 (n-mode product)**.** *For any arrays $\boldsymbol{A} \in \mathbb{R}^{q_1 \times q_2 \times \cdots \times q_d}$ and $\boldsymbol{P} \in \mathbb{R}^{p_n \times q_n}$, $n \in \{1, \ldots, d\}$, the n-mode product between $\boldsymbol{A}$ and $\boldsymbol{P}$, denoted by $\boldsymbol{A} \times_n \boldsymbol{P}$, is a array of dimension $q_1 \times q_2 \times \cdots q_{n-1} \times p_n \times q_{n+1} \times \cdots q_d$ of which $(l_1, \ldots, l_{n-1}, j, l_{n+1}, \ldots l_d)$-th element is defined by*

$$(\boldsymbol{A} \times_n \boldsymbol{P})_{l_1,\ldots,l_{n-1},j,l_{n+1},\ldots l_d} = \sum_{i=1}^{q_n} \boldsymbol{A}_{l_1,\ldots,l_{n-1},i,l_{n+1},\ldots l_d} \boldsymbol{P}_{j,i}.$$

Definition 2 (Tucker decomposition)**.** *Tucker decomposition of $\boldsymbol{A} \in \mathbb{R}^{q_1 \times q_2 \times \cdots \times q_d}$ is*

$$\boldsymbol{A} = \boldsymbol{G} \times_1 \boldsymbol{U}_1 \times_2 \cdots \times_d \boldsymbol{U}_d, \tag{3}$$

where $\boldsymbol{U}_i \in \mathbb{R}^{q_i \times r_i}$ $i = 1, 2, \ldots, d$, are called the "factor matrices" (usually orthonormal) with $r_i \leq q_i$ and $\boldsymbol{G} \in \mathbb{R}^{r_1 \times \cdots \times r_d}$ is called the "core tensor".

Figure 1 provides a pictorial illustration of a Tucker decomposition. Unlike matrices, the concept of rank is more complicated for arrays of order 3 or above. Tucker decomposition naturally leads to a particular form of rank, called "multilinear rank", which is directly related to the familiar concept of matrix ranks. To see this, we employ a reshaping operation called *matricization*, which rearranges elements of an array into a matrix.

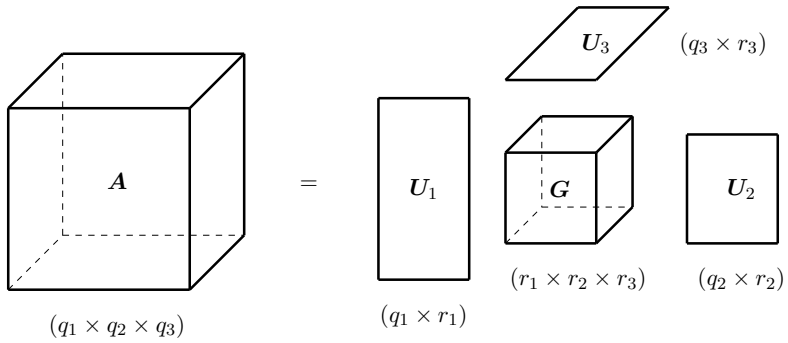


Figure 1: Tucker decomposition of a third-order array. The values in the parentheses are dimensions for the corresponding matrices or arrays.

Definition 3 (Matricization)**.** *For any $n \in \{1, \ldots, d\}$, the n-mode matricization of $\boldsymbol{A} \in \mathbb{R}^{q_1 \times q_2 \times \cdots \times q_d}$, denoted by $\boldsymbol{A}_{(n)}$, is a matrix of dimension $q_n \times (\prod_{k \neq n} q_k)$ of which (l_n, j)-th element is defined by $[\boldsymbol{A}_{(n)}]_{l_n,j} = \boldsymbol{A}_{l_1,\ldots,l_d}$, where $j = 1 + \sum_{i=1, i \neq n}^{d} (l_i - 1)(\prod_{m=1, m \neq n}^{i-1} q_m)$*[1].

[1]All empty products are defined as 1. For example, $\prod_{m=i}^{j} q_m = 1$ when $i > j$.

For any $\boldsymbol{A} \in \mathbb{R}^{q_1 \times q_2 \times \cdots \times q_d}$, by simple derivations, one can obtain a useful relationship between the n-mode matricization and Tucker decomposition $\boldsymbol{A} = \boldsymbol{G} \times_1 \boldsymbol{U}_1 \times_2 \cdots \times_d \boldsymbol{U}_d$:

$$\boldsymbol{A}_{(n)} = \boldsymbol{U}_n \boldsymbol{G}_{(n)} (\boldsymbol{U}_d \otimes \cdots \otimes \boldsymbol{U}_{n+1} \otimes \boldsymbol{U}_{n-1} \otimes \cdots \otimes \boldsymbol{U}_1)^\intercal, \tag{4}$$

where, with a slight abuse of notation, $\otimes$ also represents the Kronecker product between matrices. Hence if the factor matrices are of full rank, then $\operatorname{rank}(\boldsymbol{A}_{(n)}) = \operatorname{rank}(\boldsymbol{G}_{(n)})$. The vector $(\operatorname{rank}(\boldsymbol{A}_{(1)}), \ldots, \operatorname{rank}(\boldsymbol{A}_{(d)}))$ is known as the *multilinear rank* of $\boldsymbol{A}$. Clearly from (4), one can choose a Tucker decomposition such that $\{\boldsymbol{U}_k : k = 1, \ldots, d\}$ are orthonormal matrices and $\operatorname{rank}(\boldsymbol{U}_k) = r_k$. Therefore a "small" multilinear rank corresponds to a small core tensor and thus an intrinsic dimension reduction, which potentially improves estimation and interpretation. We will relate this low-rank structure to multidimensional functional data.

3.2 Functional unfolding for infinite-dimensional tensors

To encourage low-rank structures in covariance function estimation, we generalize the matricization operation for finite-dimensional arrays to infinite-dimensional tensors (Hackbusch, 2012). Here let $\mathcal{G} = \bigotimes_{k=1}^d \mathcal{G}_k$ denote a generic tensor product space where $\mathcal{G}_k$ is an RKHS of functions with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{G}_k}$, for $k = 1, \ldots, d$.

Notice that the tensor product space $\mathcal{G} = \bigotimes_{k=1}^d \mathcal{G}_k$ can be generated by some elementary tensors of the form $\bigotimes_{k=1}^d f_k(x_1, \ldots, x_d) = \prod_{k=1}^d f_k(x_k)$ where $f_k \in \mathcal{G}_k, k = 1, \ldots, d$. More specifically, $\mathcal{G}$ is the completion of the linear span of all elementary tensors under the inner product $\langle \bigotimes_{k=1}^d f_k, \bigotimes_{k=1}^d f_k' \rangle_{\mathcal{G}} = \prod_{k=1}^d \langle f_k, f_k' \rangle_{\mathcal{G}_k}$, for any $f_k, f_k' \in \mathcal{G}_k$.

In Definition 4 below, we generalize matricization/unfolding for finite-dimensional arrays to infinite-dimensional elementary tensors. We also define a square unfolding for infinite-dimensional tensors that will be used to describe the spectrum of covariance operators.

Definition 4 (Functional unfolding operators)**.** *The one-way unfolding operator and square unfolding operators are defined as follows for any elementary tensor of the form* $\bigotimes_{k=1}^d f_k$.

1. *One-way unfolding operator* $\mathcal{U}_j$ *for* $j = 1, \ldots, d$*: The* j*-mode one-way unfolding operator* $\mathcal{U}_j : \bigotimes_{k=1}^d \mathcal{G}_k \to \mathcal{G}_j \otimes (\bigotimes_{k \neq j} \mathcal{G}_k)$ *is defined by* $\mathcal{U}_j(\bigotimes_{k=1}^d f_k) = f_j \otimes (\bigotimes_{k \neq j} f_k)$.

2. *Square unfolding operator* $\mathcal{S}$*: When* d *is even, the square unfolding operator* $\mathcal{S} : \bigotimes_{j=1}^d \mathcal{G}_j \to (\bigotimes_{j=1}^{d/2} \mathcal{G}_j) \otimes (\bigotimes_{k=d/2+1}^d \mathcal{G}_k)$ *is defined by* $\mathcal{S}(\bigotimes_{j=1}^d f_j) = (\bigotimes_{j=1}^{d/2} f_j) \otimes (\bigotimes_{k=d/2+1}^d f_k)$.

These definitions extend to any function $f \in \mathcal{G}$ *by linearity. For notational simplicity we denote* $\mathcal{U}_j(f)$ *by* $f_{(j)}$*,* $j = 1, \ldots, d$*, and* $\mathcal{S}(f)$ *by* $f_{\blacksquare}$.

Note that the range of each functional unfolding operator, either $\mathcal{U}_j$, $j = 1, \ldots, d$ or $\mathcal{S}$, is a tensor product of *two* RKHSs, so its output can be interpreted as an (induced) operator. Given a function $f \in \mathcal{G}$, the multilinear rank can be defined as $(\operatorname{rank}(f_{(1)}), \ldots, \operatorname{rank}(f_{(d)}))$, where $f_{(j)}$'s are interpreted as an operator here and $\operatorname{rank}(A)$ is the rank of any operator A. If all $\mathcal{G}_k, k = 1, \ldots, d$ are finite-dimensional, the singular values of the output of any functional unfolding operator match with those of the j-mode matricization (of the corresponding array representation).

3.3 Functional unfolding for covariance functions

Suppose that the random field $X \in \mathcal{H} = \bigotimes_{k=1}^p \mathcal{H}_k$ where each $\mathcal{H}_k$ is a RKHS of functions equipped with an inner product $\langle \cdot, * \rangle_k$ and corresponding norm $\| \cdot \|_k$, $k = 1, \ldots, p$. Then the

covariance function γ_0 resides in $\mathcal{H} \otimes \mathcal{H} = (\bigotimes_{j=1}^{p} \mathcal{H}_j) \otimes (\bigotimes_{k=1}^{p} \mathcal{H}_k)$. To estimate γ_0, we could consider a special case of $\mathcal{G} = \bigotimes_{j=1}^{d} \mathcal{G}_j$ in Section 3.2 by letting $d = 2p$, $\mathcal{G}_j = \mathcal{H}_j$ for $j = 1, \ldots, p$; $\mathcal{G}_j = \mathcal{H}_{j-p}$ for $j = p+1, \ldots, d$; and $\langle \cdot, \cdot \rangle_{\mathcal{G}_j} = \langle \cdot, \cdot \rangle_j$ for $j = 1, \ldots, d$.

Clearly, the elements of $\mathcal{H} \otimes \mathcal{H}$ are identified by those in $\mathcal{G} = \bigotimes_{j=1}^{d} \mathcal{G}_j$. In terms of the folding structure, $\mathcal{H} \otimes \mathcal{H}$ has a squarely unfolded structure. Since a low-multilinear-rank structure is represented by different unfolded forms, it would be easier to study the completely folded space $\bigotimes_{k=1}^{d} \mathcal{G}_k$ instead of the squarely unfolded space $\mathcal{H} \otimes \mathcal{H}$. We use Γ_0 to represent the folded covariance function, the corresponding element of γ_0 in $\mathcal{G}$. In other words, $\Gamma_{0,\blacksquare} = \gamma_0$. For any $\Gamma \in \mathcal{G}$, $\operatorname{rank}(\Gamma_{\blacksquare})$ is defined as the *two-way rank* of Γ while $\operatorname{rank}(\Gamma_{(1)}), \ldots, \operatorname{rank}(\Gamma_{(p)})$ are defined as the *one-way ranks* of Γ.

Remark 1. For an array $\boldsymbol{A} \in \mathbb{R}^{\prod_{k=1}^{d} q_k}$, the one-way unfolding $\mathcal{U}_j(\boldsymbol{A})$ is the same as matricization, if we further impose the same ordering of the columns in the output of $\mathcal{U}_j(\boldsymbol{A})$, $j = 1, \ldots, d$. This ordering is just related to how we represent the array, and is not crucial in the general definition of $\mathcal{U}_j$. Since the description of the computation strategy depends on the explicit representation, we will always assume this ordering. Similarly, we also define a specific ordering of rows and columns for $\boldsymbol{A}_{\blacksquare} \in \mathbb{R}^{(d/2)\times(d/2)}$ when d is even, such that its (j_1, j_2)-th entry is $\boldsymbol{A}_{k_1,\ldots,k_d}$ where $j_1 = 1 + \sum_{i=1}^{d/2}(k_i - 1)(\prod_{m=i+1}^{d/2} q_m)$ and $j_2 = 1 + \sum_{i=d/2+1}^{d}(k_i - 1)(\prod_{m=i+1}^{d} q_m)$.

3.4 One-way and two-way ranks in covariance functions

Here we illustrate the roles of one-way and two-way ranks in the modeling of covariance functions. For a general $\mathcal{G} = \bigotimes_{j=1}^{d} \mathcal{G}_j$, let $\{e_{k,l_k} : l_j = 1, \ldots, q_k\}$ be a set of orthonormal basis functions of $\mathcal{G}_k$ for $k = 1, \ldots, d = 2p$, where q_k is allowed to be infinite, depending on the dimensionality of $\mathcal{G}_k$. Then $\{\bigotimes_{k=1}^{d} e_{k,l_k} : l_k = 1, \ldots, q_k; k = 1, \ldots, d\}$ forms a set of orthonormal basis functions for $\mathcal{G}$. Thus for any $\Gamma \in \mathcal{G}$, we can express

$$\Gamma = \sum_{k_1,k_2,\ldots,k_d} \boldsymbol{B}_{k_1,\ldots,k_d} \bigotimes_{i=1}^{d} e_{i,k_i}, \tag{5}$$

where the coefficients $\boldsymbol{B}_{k_1,\ldots,k_d}$ are real numbers. For convenience, we collectively put them into an array $\boldsymbol{B} \in \mathbb{R}^{\prod_{k=1}^{d} q_k}$.

To illustrate the low-multilinear-rank structures for covariance functions, we consider $p = 2$, i.e., $d = 2p = 4$, and then by (5) the folded covariance function Γ can be expressed by

$$\Gamma(s_1, s_2, t_1, t_2) = \sum_{k_1=1}^{q_1} \sum_{k_2=1}^{q_2} \sum_{k_3=1}^{q_1} \sum_{k_4=1}^{q_2} \boldsymbol{B}_{k_1,k_2,k_3,k_4} e_{1,k_1}(s_1) e_{2,k_2}(s_2) e_{1,k_3}(t_1) e_{2,k_4}(t_2).$$

To be precise, the covariance function is the squarely unfolded $\Gamma_{\blacksquare}((s_1, s_2), (t_1, t_2)) \equiv \Gamma(s_1, s_2, t_1, t_2)$. Suppose that $\boldsymbol{B}$ possesses (or is well-approximated by) a structure of a low multilinear rank, and yields Tucker decomposition $\boldsymbol{B} = \boldsymbol{E} \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2 \times_3 \boldsymbol{U}_1 \times_4 \boldsymbol{U}_2$[2] where $\boldsymbol{E} \in \mathbb{R}^{r_1 \times r_2 \times r_1 \times r_2}$, $\boldsymbol{U}_k \in \mathbb{R}^{q_k \times r_k}$ for $k = 1, 2$, and columns of $\boldsymbol{U}_k$ are orthonormal. Apparently $R := \operatorname{rank}(\boldsymbol{B}_{\blacksquare})$ is the two-way rank of Γ, while r_1 and r_2 are the corresponding one-way ranks. Now the covariance function can be further represented as

$$\Gamma(s_1, s_2, t_1, t_2) = \sum_{j_1=1}^{r_1} \sum_{j_2=1}^{r_2} \sum_{j_3=1}^{r_1} \sum_{j_4=1}^{r_2} \boldsymbol{E}_{j_1,j_2,j_3,j_4} u_{j_1}(s_1) v_{j_2}(s_2) u_{j_3}(t_1) v_{j_4}(t_2),$$

[2]Definition 1 is extended to the case when q_n is infinite.

where $\{u_j : j = 1, \ldots, r_1\}$ and $\{v_k : k = 1, \ldots, r_2\}$ are (possibly infinite) linear combinations of the original basis functions. In fact, $\{u_j : j = 1, \ldots, r_1\}$ and $\{v_k : k = 1, \ldots, r_2\}$ are the sets of orthonormal functions of $\mathcal{G}_1$ and $\mathcal{G}_2$ respectively. Apparently $\text{rank}(\boldsymbol{E}_{\blacksquare}) = R$.

Consider the eigen-decomposition of the squarely unfolded core tensor $\boldsymbol{E}_{\blacksquare} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^T$ where $\boldsymbol{D} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_R)$ and $\boldsymbol{P} \in \mathbb{R}^{r_1 r_2 \times R}$ has orthonormal columns. Then we obtain the eigen-decomposition of the covariance function $\Gamma_{\blacksquare}$:

$$\Gamma_{\blacksquare}((s_1, s_2), (t_1, t_2)) = \sum_{g=1}^{R} \lambda_g f_g(s_1, s_2) f_g(t_1, t_2),$$

where the eigenfunction is

$$f_g(s_1, s_2) = \sum_{j_1=1}^{r_1} \sum_{j_2=1}^{r_2} \boldsymbol{P}_{j_2+(j_1-1)r_1, g} u_{j_1}(s_1) v_{j_2}(s_2) =: \begin{cases} \sum_{j_1=1}^{r_1} a_{j_1,g}(s_2) u_{j_1}(s_1) \\ \sum_{j_2=1}^{r_2} b_{j_2,g}(s_1) v_{j_2}(s_2) \end{cases},$$

with $a_{j_1,g}(\cdot) = \sum_{j_2=1}^{r_2} \boldsymbol{P}_{j_2+(j_1-1)r_1,g} v_{j_2}(\cdot)$ and $b_{j_2,g}(\cdot) = \sum_{j_1=1}^{r_1} \boldsymbol{P}_{j_2+(j_1-1)r_1,g} u_{j_1}(\cdot)$.

First, this indicates that the two-way rank R is the same as the rank of the covariance operator. Second, this shows that $\{u_{j_1} : j_1 = 1, \ldots, r_1\}$ is the common basis for the variation along the dimension s_1, hence describing the marginal structure along s_1. Similarly $\{v_{j_2} : j_2 = 1, \ldots, r_2\}$ is the common basis that characterizes the marginal variation along the dimension s_2. We call them the *marginal basis* along the respective dimension. Therefore, the one-way ranks r_1 and r_2 are the minimal numbers of the one-dimensional functions for the dimensions s_1 and s_2 respectively that construct all the eigenfunctions of covariance function Γ.

Similarly, for p-dimensional functional data, each eigenfunction can be represented by a linear combination of p-products of univariate functions. One can then show that the two-way rank R is the same as the rank of the covariance operator and the one-way ranks $r_1, \ldots, r_p$ are the minimal numbers of one-dimensional functions along respective dimensions that characterize all eigenfunctions of the covariance operator.

Remark 2. Obviously, $R \leq \prod_{k=1}^{p} r_k$ for p-dimensional functional data. If the random field X has the property of "weak separability" as defined by Lynch and Chen (2018), then $\max(r_1, \ldots, r_p) \leq R$ so the low-rank structure in terms of R will be automatically translated to low one-way ranks. Note that the construction of our estimator and corresponding theoretical analysis *do not* require separability conditions.

Compared to typical low-rank covariance modelings only in terms of R, we also intend to regularize the one-way ranks $r_1, \ldots, r_p$ for two reasons. First, the illustration above shows that the structure of low one-way ranks encourages a "sharing" structure of one-dimensional variations among different eigenfunctions. Promoting low one-way ranks can facilitate additional dimension reduction and further alleviates the curse of dimensionality. Moreover, one-dimensional marginal structures will provide more details of the covariance function structure and thus help with a better understanding of p-dimensional eigenfunctions.

Therefore, we will utilize both one-way and two-way structures and propose an estimation procedure that regularizes one-way and two-way ranks jointly and flexibly, with the aim of seeking the "sharing" of marginal structures while controlling the number of eigen-components simultaneously.

4 Covariance Function Estimation

In this section we propose a low-rank covariance function estimation framework based on functional unfolding operators and spectral regularizations. Spectral penalty functions (Abernethy et al., 2009; Wong and Zhang, 2019) are defined as follows.

Definition 5 (Spectral penalty function)**.** *Given a compact operator A, a spectral penalty function takes the form $\Psi(A) = \sum_{k\geq 1} \psi(\lambda_k(A))$ with the singular values of the operator A, $\lambda_1(A)$, $\lambda_2(A)$, ... in a descending order of magnitude and a non-decreasing penalty function ψ such that $\psi(0) = 0$.*

Recall $\mathcal{H} = \bigotimes_{j=1}^{p} \mathcal{H}_j$ and $\mathcal{G} = \bigotimes_{j=1}^{d} \mathcal{G}_j$ where $d = 2p$, $\mathcal{G}_j = \mathcal{H}_j$ for $j = 1, \ldots, p$, and $\mathcal{G}_j = \mathcal{H}_{j-p}$ for $j = p+1, \ldots, d$. Clearly, a covariance operator is self-adjoint and positive semi-definite. Therefore we consider the hypothesis space $\mathcal{M}^+ = \{\Gamma \in \mathcal{M} : \langle \Gamma_{\blacksquare} f, f\rangle_{\mathcal{H}} \geq 0, \text{for all } f \in \mathcal{H}\}$, where $\mathcal{M} = \{\Gamma \in \mathcal{G} : \Gamma_{\blacksquare} \text{ is self-adjoint}\}$, and propose a general class of covariance function estimators as follows:

$$\underset{\Gamma \in \mathcal{M}^+}{\arg\min} \left\{ \ell(\Gamma) + \lambda \left[\beta \Psi_0(\Gamma_{\blacksquare}) + \frac{1-\beta}{p} \sum_{j=1}^{p} \Psi_j(\Gamma_{(j)}) \right] \right\}, \tag{6}$$

where ℓ is a convex and smooth loss function, $\{\Psi_j : j = 1, \ldots, p\}$ are spectral penalty functions, and $\lambda \geq 0$, $\beta \in [0, 1]$ are tuning parameters. Here Ψ_0 penalizes the squarely unfolded operator $\Gamma_{\blacksquare}$ while Ψ_j regularizes one-way unfolded operator $\Gamma_{(j)}$ respectively for $j = 1, \ldots, p$. The tuning parameter β controls the relative degree of regularization between one-way and two-way singular values. The larger the β is, the more penalty is imposed on the two-way singular values relative to the one-way singular values. When $\beta = 1$, the penalization is only on the eigenvalues of the covariance operator (i.e., the two-way singular values), similarly as Wong and Zhang (2019).

To achieve low-rank estimation, we adopt a special form of (6):

$$\hat{\Gamma} = \underset{\Gamma \in \mathcal{M}^+}{\arg\min} \left\{ \ell_{\text{square}}(\Gamma) + \lambda \left[\beta \|\Gamma_{\blacksquare}\|_* + \frac{1-\beta}{p} \sum_{j=1}^{p} \|\Gamma_{(j)}\|_* \right] \right\}, \tag{7}$$

where $\|\cdot\|_*$ is the sum of singular values, also called trace norm, and ℓ_{square} is the squared error loss:

$$\ell_{\text{square}}(\Gamma) = \frac{1}{nm(m-1)} \sum_{i=1}^{n} \sum_{1 \leq j \neq j' \leq m} \{\Gamma(T_{ij1}, \ldots, T_{ijp}, T_{ij'1}, \ldots, T_{ij'p}) - Z_{ijj'}\}^2, \tag{8}$$

with $Z_{ijj'} = \{Y_{ij} - \hat{\mu}(T_{ij1}, \ldots, T_{ijp})\}\{Y_{ij'} - \hat{\mu}(T_{ij'1}, \ldots, T_{ij'p})\}$, $\hat{\mu}$ as an estimate of the mean function, and T_{ijk} as the k-th element of location vector $\boldsymbol{T}_{ij}$. Notice that trace-norm regularizations promote low-rankness of the underlying operators, hence leading to a low-rank estimation in terms of both the one-way and two-way (covariance) ranks.

4.1 Representer theorem and parametrization

Before deriving a computational algorithm, we notice that the optimization (7) is an infinite-dimensional optimization which is generally unsolvable. To overcome this challenge, we show that the solution to (7) always lies in a known finite-dimensional sub-space given data, hence allowing a finite-dimensional parametrization. Indeed, we are able to achieve a stronger result in Theorem 1 which holds for the general class of estimators obtained by (6).

Let $\mathcal{L}_{n,m} = \{T_{ijk} : i = 1, \ldots, n, j = 1, \ldots, m, k = 1, \ldots, p\}$.

Theorem 1 (Representer theorem)**.** *If the solution set of* (6) *is not empty, there always exists a solution* Γ *lying in the space* $\mathcal{G}(\mathcal{L}_{n,m}) := \bigotimes_{k=1}^{2p} \mathcal{K}_k$*, where* $\mathcal{K}_{p+k} = \mathcal{K}_k$ *and* $\mathcal{K}_k = \operatorname{span}\{K_k(T_{ijk}) : i = 1, \ldots, n, j = 1, \ldots, m\}$ *for* $k = 1, \ldots, p$*. The solution takes the form:*

$$\Gamma(s_1, \ldots, s_p, t_1, \ldots, t_p) = \boldsymbol{A} \times_1 \boldsymbol{z}_1^\intercal(s_1) \times_2 \boldsymbol{z}_2^\intercal(s_2) \cdots \times_p \boldsymbol{z}_p^\intercal(s_p) \times_{p+1} \boldsymbol{z}_1^\intercal(t_1) \cdots \times_{2p} \boldsymbol{z}_p^\intercal(t_p), \tag{9}$$

where the l*-th element of* $\boldsymbol{z}_k(\cdot) \in \mathbb{R}^{mn}$ *is* $K(T_{ijk}, \cdot)$ *with* $l = (i-1)n + j$*. Also,* $\boldsymbol{A}$ *is a* $2p$*-th order tensor where the dimension of each mode is* nm *and* $\boldsymbol{A}_\blacksquare$ *is a symmetric matrix.*

The proof of Theorem 1 is given in Section S1 of the supplementary material. By Theorem 1, we can now only focus on covariance function estimators of the form (9). Let $\boldsymbol{B} = \boldsymbol{A} \times_1 \boldsymbol{M}_1^T \cdots \times_p \boldsymbol{M}_p^T \times_{p+1} \boldsymbol{M}_1^T \cdots \times_{2p} \boldsymbol{M}_p^T$, where $\boldsymbol{M}_k$ is a $nm \times q_k$ matrix such that $\boldsymbol{M}_k \boldsymbol{M}_k^T = \boldsymbol{K}_k = [K(T_{i_1,j_1,k}, T_{i_2,j_2,k})]_{1 \le i_1, i_2 \le n, 1 \le j_1, j_2 \le m}$. With $\boldsymbol{B}$, we can express

$$\begin{aligned} \Gamma(s_1, \ldots, s_p, t_1, \ldots, t_p) = \boldsymbol{B} \times_1 \{\boldsymbol{M}_1^+ \boldsymbol{z}_1(s_1)\}^\intercal \cdots \times_p \{\boldsymbol{M}_p^+ \boldsymbol{z}_p(s_p)\}^\intercal \\ \times_{p+1} \{\boldsymbol{M}_1^+ \boldsymbol{z}_1(t_1)\}^\intercal \cdots \times_{2p} \{\boldsymbol{M}_p^+ \boldsymbol{z}_p(t_p)\}^\intercal, \end{aligned} \tag{10}$$

where $z_k(\cdot)$ is defined in Theorem 1 and $\boldsymbol{M}_k^+$ is the MoorePenrose inverse of matrix $\boldsymbol{M}_k$.

The Gram matrix $\boldsymbol{K}_k$ is often approximately low-rank. For computational simplicity, one could adopt q_k to be significantly smaller than nm. Ideally we can obtain the "best" low-rank approximation with respect to the Frobenius norm by eigen-decomposition, but a full eigen-decomposition is computationally expensive. Instead, randomized algorithms can be used to obtain low-rank approximations in an efficient manner (Halko et al., 2009).

One can easily show that the eigenvalues of the operator $\Gamma_\blacksquare$ are the same as those of the matrix $\boldsymbol{B}_\blacksquare$ and that the singular values of the operator $\Gamma_{(j)}$ are the same as those of the matrix $\boldsymbol{B}_{(j)}$. Therefore, solving (7) is equivalent to solving the following optimization:

$$\min_{\boldsymbol{B}} \left\{ \tilde{\ell}_{\text{square}}(\boldsymbol{B}) + \lambda \left[\beta h(\boldsymbol{B}_\blacksquare) + \frac{1-\beta}{p} \sum_{k=1}^{p} \left\| \boldsymbol{B}_{(j)} \right\|_* \right] \right\}, \tag{11}$$

where $\|\cdot\|_*$ also represents the trace norm of matrices, $h(\boldsymbol{H}) = \|\boldsymbol{H}\|_*$ if matrix $\boldsymbol{H}$ is positive semi-definite, and $h(\boldsymbol{H}) = \infty$ otherwise, and $\tilde{\ell}_{\text{square}}(\boldsymbol{B}) = \ell_{\text{square}}(\Gamma)$, where Γ is constructed from (10).

Beyond estimating the covariance function, one may be further interested in the eigen-decomposition of $\Gamma_\blacksquare$ via the L^2 inner product, e.g., to perform functional principal component analysis in the usual sense. Due to the finite-dimensional parametrization, a closed-form expression of L^2 eigen-decomposition can be derived from our estimator without further discretization or approximation. In addition, we can obtain a similar one-way analysis in terms of the L_2 inner product. We can define a L^2 singular value decomposition via the Tucker form and obtain the L^2 marginal basis. Details are given in Appendix A.

4.2 Computational algorithm

We solve (11) by the accelerated alternating direction method of multipliers (ADMM) algorithm (Kadkhodaie et al., 2015). We begin with an alternative form of (11):

$$\min_{\boldsymbol{B} \in \mathbb{R}^{q_1 \times \cdots \times q_{2p}}} \left\{ \tilde{\ell}_{\text{square}}(\boldsymbol{B}) + \lambda\beta h(\boldsymbol{D}_{0,\blacksquare}) + \lambda \frac{1-\beta}{p} \sum_{k=1}^{p} \left\| \boldsymbol{D}_{j,(j)} \right\|_* \right\}. \tag{12}$$

$$\text{subject to} \quad \boldsymbol{B} = \boldsymbol{D}_0 = \boldsymbol{D}_1 = \cdots = \boldsymbol{D}_p \tag{13}$$

where $q_{p+k} = q_k$ for $k = 1, \ldots, p$.

Then a standard ADMM algorithm solves the optimization problem (12) by minimizing the augmented Lagrangian with respect to different variables alternatively. More explicitly, at the $(t+1)$-th iteration, the following updates are implemented:

$$\boldsymbol{B}^{(t+1)} = \underset{\boldsymbol{B}}{\operatorname{argmin}} \left\{ \tilde{\ell}_{\text{square}}(\boldsymbol{B}) + \frac{\eta}{2} \|\boldsymbol{B}_{\blacksquare} - \boldsymbol{D}_{0,\blacksquare}^{(t)} + \boldsymbol{V}_{0,\blacksquare}^{(t)}\|_F^2 + \frac{\eta}{2} \sum_{k=1}^{p} \left\| \boldsymbol{B}_{(k)} - \boldsymbol{D}_{k,(k)}^{(t)} + \boldsymbol{V}_{k,(k)}^{(t)} \right\|_F^2 \right\}, \tag{14a}$$

$$\boldsymbol{D}_0^{(t+1)} = \underset{\boldsymbol{D}_0}{\operatorname{argmin}} \left\{ \lambda\beta h(\boldsymbol{D}_{0,\blacksquare}) + \frac{\eta}{2} \left\| \boldsymbol{B}_{\blacksquare}^{(t+1)} - \boldsymbol{D}_{0,\blacksquare} + \boldsymbol{V}_{0,\blacksquare}^{(t)} \right\|_F^2 \right\}, \tag{14b}$$

$$\boldsymbol{D}_k^{(t+1)} = \underset{\boldsymbol{D}_k}{\operatorname{argmin}} \left\{ \lambda \frac{1-\beta}{p} \|\boldsymbol{D}_{k,(k)}\|_* + \frac{\eta}{2} \left\| \boldsymbol{B}_{(k)}^{(t+1)} - \boldsymbol{D}_{k,(k)} + \boldsymbol{V}_{k,(k)}^{(t)} \right\|_F^2 \right\}, \ k = 1, \ldots, p, \tag{14c}$$

$$\boldsymbol{V}_k^{(t+1)} = \boldsymbol{V}_k^{(t)} + \boldsymbol{B}^{(t+1)} - \boldsymbol{D}_k^{(t+1)}, \ k = 0, \ldots, p, \tag{14d}$$

where $\boldsymbol{V}_k \in \mathbb{R}^{q_1 \times \cdots q_{2p}}$, for $k = 0, \ldots, p$, are scaled Lagrangian multipliers and $\eta > 0$ is an algorithmic parameter. An adaptive strategy to tune η is provided in Boyd et al. (2010). One can see that Steps (14a), (14b) and (14c) involve additional optimizations. Now we discuss how to solve them.

The objective function of (14a) is a quadratic function, and so we can easily solve this with a closed-form solution, given in line 2 of Algorithm 1. To solve (14b) and (14c), we use proximal operator prox_v^k, $k = 1, \ldots, p$ and $\text{prox}_v^+ : \mathbb{R}^{q_1 \times \cdots \times q_{2p}} \to \mathbb{R}^{q_1 \times \cdots \times q_{2p}}$ respectively defined by

$$\text{prox}_v^k(\boldsymbol{A}) = \underset{\boldsymbol{W} \in \mathbb{R}^{q_1 \times \cdots \times q_{2p}}}{\operatorname{argmin}} \left\{ \frac{1}{2} \|\boldsymbol{W}_{(k)} - \boldsymbol{A}_{(k)}\|_F^2 + v\|\boldsymbol{W}_{(k)}\|_* \right\}, \tag{15a}$$

$$\text{prox}_v^+(\boldsymbol{A}) = \underset{\boldsymbol{W} \in \mathbb{R}^{q_1 \times \cdots \times q_{2p}}}{\operatorname{argmin}} \left\{ \frac{1}{2} \|\boldsymbol{W}_{\blacksquare} - \boldsymbol{A}_{\blacksquare}\|_F^2 + v h(\boldsymbol{W}_{\blacksquare}) \right\}, \tag{15b}$$

for $v \geq 0$. By Lemma 1 in Mazumder et al. (2010), the solutions to (15) have closed forms.

For (15a), write the singular value decomposition of $\boldsymbol{A}_{(k)}$ as $\boldsymbol{U}\text{diag}((\tilde{a}_1, \ldots, \tilde{a}_{q_k}))\boldsymbol{V}^\intercal$, then $[\text{prox}_v^k(\boldsymbol{A})]_{(k)} = \boldsymbol{U}\text{diag}(\tilde{\boldsymbol{c}})\boldsymbol{V}^\intercal$ where $\tilde{\boldsymbol{c}} = ((\tilde{a}_1 - v)_+, (\tilde{a}_2 - v)_+, \ldots, (\tilde{a}_{q_k} - v)_+)$. As for (15b), is restricted to be a symmetric matrix since the penalty h equals infinity otherwise. Thus (15b) is equivalent to minimizing $\left\{(1/2)\|\boldsymbol{W}_{\blacksquare} - (\boldsymbol{A}_{\blacksquare} + \boldsymbol{A}_{\blacksquare}^\intercal)/2\|_F^2 + v h(\boldsymbol{W}_{\blacksquare})\right\}$ since $\langle \boldsymbol{W}_{\blacksquare}, (\boldsymbol{A}_{\blacksquare} - \boldsymbol{A}_{\blacksquare}^\intercal)/2 \rangle = \langle (\boldsymbol{W}_{\blacksquare} + \boldsymbol{W}_{\blacksquare}^\intercal)/2, (\boldsymbol{A}_{\blacksquare} - \boldsymbol{A}_{\blacksquare}^\intercal)/2 \rangle = 0$. Suppose that $(\boldsymbol{A}_{\blacksquare} + \boldsymbol{A}_{\blacksquare}^\intercal)/2$ yields eigen-decomposition $\boldsymbol{P}\text{diag}((\tilde{a}_1, \ldots, \tilde{a}_q)\boldsymbol{P}^\intercal$. Then $[\text{prox}_v^+(\boldsymbol{A})]_{\blacksquare} = \boldsymbol{P}\text{diag}(\tilde{\boldsymbol{c}})\boldsymbol{P}^\intercal$, where $\tilde{\boldsymbol{c}} = ((\tilde{a}_1 - v)_+, (\tilde{a}_2 - v)_+, \ldots, (\tilde{a}_q - v)_+)$. Unlike singular values, the eigenvalues may be negative. Hence, as opposed to prox_v^k, this procedure prox_v^+ also removes eigen-components with negative eigenvalues.

The details of computational algorithm are given in Algorithm 1, an accelerated version of ADMM which involves additional steps for a faster algorithmic convergence.

5 Asymptotic Properties

In this section, we conduct an asymptotic analysis for the proposed estimator $\hat{\Gamma}$ as defined in (7). Our analysis has a unified flavor such that the derived convergence rate of the proposed estimator automatically adapts to sparse and dense settings. Throughout this section, we neglect the mean function estimation error by setting $\mu_0(\boldsymbol{t}) = \hat{\mu}(\boldsymbol{t}) = 0$ for any $\boldsymbol{t} \in \mathcal{T}$, which leads to a cleaner and more focused analysis. The additional error from the mean function estimation can be incorporated into our proofs without any fundamental difficulty.

5.1 Assumptions

Without loss of generality let $\mathcal{T} = [0,1]^p$. The assumptions needed in the asymptotic results are listed as follows.

Assumption 1. *Sample fields* $\{X_i : i = 1, \ldots, n\}$ *reside in* $\mathcal{H} = \bigotimes_{k=1}^{p} \mathcal{H}_k$ *where* $\mathcal{H}_k$ *is an RKHS of functions on* $[0,1]$ *with a continuous and square integrable reproducing kernel* K_k.

Assumption 2. *The true (folded) covariance function* $\Gamma_0 \neq 0$ *and* $\Gamma_0 \in \mathcal{G} = \bigotimes_{j=1}^{d} \mathcal{G}_j$, *where* $d = 2p$, $\mathcal{G}_j = \mathcal{H}_j$ *for* $j = 1, \ldots, p$ *and* $\mathcal{G}_j = \mathcal{H}_{j-p}$ *for* $j = p+1, \ldots, d$.

Assumption 3. *The locations* $\{\boldsymbol{T}_{ij} : i = 1, \ldots, n; j = 1, \ldots, m\}$ *are independent random vectors from* Uniform$[0,1]^p$, *and they are independent of* $\{X_i : i = 1, \ldots, n\}$.
The errors $\{\epsilon_{ij} : i = 1, \ldots, n; j = 1, \ldots, m\}$ *are independent of both locations and sample fields.*

Assumption 4. *For each* $\boldsymbol{t} \in \mathcal{T}$, $X(\boldsymbol{t})$ *is sub-Gaussian with a parameter* $b_X > 0$ *which does not depend on* $\boldsymbol{t}$, *i.e.,* $\mathbb{E}[\exp\{\beta X(\boldsymbol{t})\}] \leq \exp\{b_X^2 \beta^2 / 2\}$ *for all* β *and* $\boldsymbol{t} \in \mathcal{T}$.

Assumption 5. *For each* i *and* j, ϵ_{ij} *is a mean-zero sub-Gaussian random variable with a parameter* b_ϵ *independent of* i *and* j, *i.e.,* $\mathbb{E}[\exp\{\beta \epsilon_{ij}\}] \leq \exp\{b_\epsilon^2 \beta^2 / 2\}$.
Moreover all errors $\{\epsilon_{ij} : i = 1, \ldots, n; j = 1, \ldots, m\}$ *are independent.*

Assumption 1 delineates a tensor product RKHS modeling, commonly seen in the nonparametric regression literature (e.g., Wahba, 1990; Gu, 2013). In Assumption 2, the condition $\Gamma_0 \in \mathcal{G}$ is satisfied if $\mathbb{E}\|X\|_{\mathcal{H}}^2 < \infty$, as shown in Cai and Yuan (2010). Assumption 3 is specified for random design and we adopt the uniform distribution here for simplicity. The uniform distribution on $[0,1]^p$ can be generalized to any other continuous distribution of which density function π satisfies $c_\pi \leq \pi(\boldsymbol{t}) \leq c'_\pi$ for all $\boldsymbol{t} \in [0,1]^p$, for some constants $0 < c_\pi \leq c'_\pi < 1$, to ensure both Theorems 2 and 3 still hold. Assumptions 4 and 5 involve sub-Gaussian conditions of the stochastic process and measurement error, which are standard tail conditions.

5.2 Reproducing kernels

In Assumption 1, the "smoothness" of the function in the underlying RKHS is not explicitly specified. It is well-known that such smoothness conditions are directly related to the eigen-decay of the respective reproducing kernel. By Mercer's Theorem (Mercer, 1909), the reproducing kernel $K_{\mathcal{H}}((t_1, \ldots, t_p), (t'_1, \ldots, t'_p))$ of $\mathcal{H}$ possesses the eigen-decomposition

$$K_{\mathcal{H}}((t_1, \ldots, t_p), (t'_1, \ldots, t'_p)) = \sum_{l=1}^{\infty} \mu_l \phi_l(t_1, \ldots, t_p) \phi_l(t'_1, \ldots, t'_p), \tag{16}$$

where $\{\mu_l : l \geq 1\}$ are non-negative eigenvalues and $\{\phi_l : l \geq 1\}$ are L^2 eigenfunctions on $[0,1]^p$. Then for the space $\mathcal{H} \otimes \mathcal{H}$, which is also identified by $\mathcal{G} = \bigotimes_{k=1}^{d} \mathcal{G}_k$, its corresponding reproducing kernel $K_{\mathcal{G}}$ has the following eigen-decomposition

$$\begin{aligned}
&K_{\mathcal{G}}((x_1, \ldots, x_{2p}), (x'_1, \ldots, x'_{2p})) \\
&= K_{\mathcal{H}}((x_1, \ldots, x_p), (x'_1, \ldots, x'_p)) K_{\mathcal{H}}((x_{p+1}, \ldots, x_{2p}), (x'_{p+1}, \ldots, x'_{2p})) \\
&= \sum_{l,h=1}^{\infty} \mu_l \mu_h \phi_l(x_1, \ldots, x_p) \phi_h(x_{p+1}, \ldots, x_{2p}) \phi_l(x'_1, \ldots, x'_p) \phi_h(x'_{p+1}, \ldots, x'_{2p}),
\end{aligned}$$

where $\{\mu_l\mu_h : l, h \geq 1\}$ are the eigenvalues of $K_{\mathcal{G}}$. Due to continuity assumption (Assumption 1) of the univariate kernels, there exists a constant b such that

$$\sup_{(x_1,\ldots,x_{2p})\in[0,1]^{2p}} K_{\mathcal{G}}((x_1,\ldots,x_{2p}),(x_1,\ldots,x_{2p})) \leq b.$$

The decay rate of the eigenvalues $\{\mu_l\mu_h : l, h \geq 1\}$ is involved in our analysis through two quantities $\kappa_{n,m}$ and $\eta_{n,m}$, which have relatively complex forms defined in Appendix B. Similar quantities can be found in other analyses of RKHS-based estimators (e.g., Raskutti et al., 2012) that accommodate general choices of RKHS. Generally $\kappa_{n,m}$ and $\eta_{n,m}$ are expected to diminish in certain orders of n and m, characterized by the decay rate of the eigenvalues $\{\mu_l\mu_h\}$. The smoother the functions in the RKHS, the faster these two quantities diminish. Our general results in Theorems 2 and 3 are specified in terms of these quantities. To provide a solid example, we derive the orders of $\kappa_{n,m}$ and $\eta_{n,m}$ under a Sobolev-Hilbert space setting and provide the convergence rate of the proposed estimator in Corollary 1.

5.3 Unified rates of convergence

We write the penalty in (7) as $I(\Gamma) = \beta\|\Gamma_{\blacksquare}\|_* + (1-\beta)p^{-1}\sum_{k=1}^{p}\|\Gamma_{(k)}\|_*$. For arbitrary functions $g_1, g_2 \in \mathcal{G}$, define their empirical inner product and the corresponding (squared) empirical norm as

$$\langle g_1, g_2\rangle_{n,m} = \frac{1}{nm(m-1)}\sum_{i=1}^{n}\sum_{1\leq j,j'\leq m} g_1(T_{ij1},\ldots,T_{ijp},T_{ij'1},\ldots,T_{ij'p})g_2(T_{ij1},\ldots,T_{ijp},T_{ij'1},\ldots,T_{ij'p}),$$

$$\|g_1\|_{n,m}^2 = \langle g_1, g_1\rangle_{n,m}.$$

Additionally, the L^2 norm of a function g is defined as $\|g\|_2 = \{\int_{\mathcal{T}} g^2(\boldsymbol{t})\,d\boldsymbol{t}\}^{1/2}$.

Define $\xi_{n,m} = \max\{\eta_{n,m}, \kappa_{n,m}, (n^{-1}\log n)^{1/2}\}$. We first provide the empirical L^2 rate of convergence for $\hat{\Gamma}$.

Theorem 2. *Suppose that Assumptions 1–5 hold. Assume $\xi_{n,m}$ satisfies $(\log n)/n \leq \xi_{n,m}^2/(\log\log\xi_{n,m}^{-1})$. If $\lambda \geq L_1\xi_{n,m}^2$ for some constant $L_1 > 0$ depending on b_X, b_ϵ and b, we have*

$$\|\hat{\Gamma} - \Gamma_0\|_{n,m} \leq \sqrt{2I(\Gamma_0)\lambda} + L_1\xi_{n,m},$$

with probability at least $1 - \exp(-cn\xi_{n,m}^2/\log n)$ for some positive universal constant c.

Next, we provide the L^2 rate of convergence for $\hat{\Gamma}$.

Theorem 3. *Under the same conditions as Theorem 2, there exist a positive constant L_2 depending on b_X, b_ϵ, b and $I(\Gamma_0)$, such that*

$$\|\hat{\Gamma} - \Gamma_0\|_2 \leq 2\sqrt{I(\Gamma_0)\lambda} + L_2\xi_{n,m},$$

with probability at least $1 - \exp(-c_p n\xi_{n,m}^2/\log n)$ for some constant c_p depending on b.

The proofs of Theorems 2 and 3 can be found in Section S1 in the supplementary material.

Remark 3. Theorems 2 and 3 are applicable to general RKHS $\mathcal{H}$ which satisfies Assumption 1. The convergence rate depends on the eigen-decay rates of the reproducing kernel. A special case of polynomial decay rates for univariate RKHS will be given in Corollary 1. Moreover, our analysis has a *unified* flavor in the sense that the resulting convergence rates automatically adapt to the orders of both n and m. In Remark 5 we will provide a discussion of a "phase transition" between dense and sparse functional data revealed by our theory.

Remark 4. With a properly chosen λ, Theorems 2 and 3 bound the convergence rates (in terms of both the empirical and theoretical L^2 norm) by $\xi_{n,m}$, which cannot be faster than $(n^{-1}\log n)^{1/2}$. The logarithmic order is due to the use of Adamczak bound in Lemma S2 in the supplementary material. If one further assumes boundedness for the sample fields X_i's (in terms of the sup-norm) and the noise variables ϵ_{ij}'s, we can instead use Talagrand concentration inequality (Bousquet bound in Koltchinskii (2011)) and the results in Theorems 2 and 3 can be improved to $\max\{\|\hat{\Gamma} - \Gamma_0\|_{n,m}^2, \|\hat{\Gamma} - \Gamma_0\|_2^2\} = \mathcal{O}_{\mathrm{p}}(\tilde{\xi}_{n,m}^2)$, where $\tilde{\xi}_{n,m} = \max\{\eta_{n,m}, \kappa_{n,m}, n^{-1/2}\}$.

Next we focus on a special case where the reproducing kernels of the univariate RKHS $\mathcal{H}_k$'s exhibit polynomial eigen-decay rates, which holds for a range of commonly used RKHS. A canonical example is α-th order Sobolev-Hilbert space:

$$\mathcal{H}_k = \{f : f^{(r)}, r = 0, \ldots, \alpha, \text{are absolutely continuous}; f^{(\alpha)} \in L^2([0,1])\},$$

where $k = 1, \ldots, p$. Here α is the same as α in Corollary 1. To derive the convergence rates, we relate the eigenvalues ν_l in (16) to the univariate RKHS $\mathcal{H}_k$, $k = 1, \ldots, p$. Due to Mercer's Theorem, the reproducing kernel K_k of $\mathcal{H}_k$ yields an eigen-decomposition with non-negative eigenvalues $\{\mu_l^{(k)} : l \geq 1\}$ and an L^2 eigenfunction $\{\phi_l^{(k)} : l \geq 1\}$, i.e., $K_k(t,t') = \sum_{l=1}^{\infty} \mu_l^{(k)} \phi_l^{(k)}(t)\phi_l^{(k)}(t')$. Therefore, the set of eigenvalues $\{\mu_l : l \geq 1\}$ in (16) is the same as the set $\{\prod_{k=1}^{p} \mu_{l_k}^{(k)} : l_1, \ldots, l_p \geq 1\}$. Given the eigen-decay of $\mu_l^{(k)}$, one can obtain the order of $\xi_{n,m}$ and hence the convergence rates from Theorems 2 and 3. Here are the results under the setting of a polynomial eigen-decay.

Corollary 1. *Suppose that the same conditions in Theorem 3 hold. If the eigenvalues of K_k for $\mathcal{H}_k, k = 1, \ldots, p$, have polynomial decaying rates, that is, there exists $\alpha > 1/2$ such that $\mu_l^{(k)} \asymp l^{-2\alpha}$ for all $k = 1, \ldots, p$, then*

$$\max\left\{\|\hat{\Gamma} - \Gamma_0\|_{n,m}^2, \|\hat{\Gamma} - \Gamma_0\|_2^2\right\} = \mathcal{O}_{\mathrm{p}}\left(\max\left\{(nm)^{-\frac{2\alpha}{1+2\alpha}}\{\log(nm)\}^{\frac{2\alpha(2p-1)}{2\alpha+1}}, \frac{\log n}{n}\right\}\right).$$

Remark 5. All Theorems 2 and 3 and Corollary 1 reveal a "phase-transition" of the convergence rate depending on the relative magnitudes between n, the sample size, and m, the number of observations per field. When $\kappa_{n,m}^2 \ll (\log n/n)$, which is equivalent to $m \gg n^{1/(2\alpha)}(\log n)^{2p-2-1/(2\alpha)}$ in Corollary 1, both empirical and theoretical L^2 rates of convergence can achieve the near-optimal rate $\sqrt{\log n/n}$. Under the stronger assumptions in Remark 4, the convergence rate will achieve the optimal order $\sqrt{1/n}$ when $\kappa_{n,m}^2 \ll 1/n$ (or $m \gg n^{1/(2\alpha)}(\log n)^{2p-1}$ in Corollary 1). In this case, the observations are so densely sampled that we can estimate the covariance function as precisely as if the entire sample fields are observable. On the contrary, when $\kappa_{n,m}^2 \gg (\log n/n)$ (or $m \ll n^{1/(2\alpha)}(\log n)^{2p-2-1/(2\alpha)}$ in Corollary 1), the convergence rate is determined by the total number of observations nm. When $p = 1$, the asymptotic result in Corollary 1, up to some $\log m$ and $\log n$ terms, is the same as the minimax optimal rate obtained by Cai and Yuan (2010), and is comparable to the L^2 rate obtained by Paul and Peng (2009) for $\alpha = 2$.

Remark 6. For covariance function estimation for unidimensional functional data, i.e., $p = 1$, a limited number of approaches, including Cai and Yuan (2010), Li and Hsing (2010), Zhang and Wang (2016), and Liebl (2019), can achieve unified theoretical results in the sense that they hold for all relative magnitudes of n and m. The similarity of these approaches is the availability of a closed form for each covariance function estimator. In contrast, our estimator obtained from (7) does not

have a closed form due to the non-differentiability of the penalty, but it can still achieve unified theoretical results which hold for both unidimensional and multidimensional functional data. Due to the lack of a closed form of our covariance estimator, we used the empirical process techniques (e.g., Bartlett et al., 2005; Koltchinskii, 2011) in the theoretical development. In particular, we have developed a novel grouping lemma (Lemma S4 in the supplementary material) to *deterministically* decouple the dependence within a U-statistics of order 2. We believe this lemma is of independent interest. In our analysis, the corresponding U-statistics is *indexed* by a function class, and this grouping lemma provides a tool to obtain uniform results (see Lemma S3 in the supplementary material). In particular, this allows us to relate the empirical and theoretical L^2 norm of the underlying function class, in precise enough order dependence on n and m to derive the unified theory. See Lemma S3 for more details. To the best of our knowledge, this paper is one of the first in the FDA literature that derives a unified result in terms of empirical process theories, and the proof technique is potentially useful for some other estimators without a closed form.

6 Simulation

To evaluate the practical performance of the proposed method, we conducted a simulation study. We in particular focused on two-dimensional functional data. Let $\mathcal{H}_1$ and $\mathcal{H}_2$ both be the RKHS with kernel $K(t_1, t_2) = \sum_{k=1}^{\infty} (k\pi)^{-4} e_k(t_1) e_k(t_2)$, where $e_k(t) = \sqrt{2}\cos(k\pi t)$, $k \geq 1$. This RKHS has been used in various studies in FDA, e.g., the simulation study of Cai and Yuan (2012). Each X_i is generated from a mean-zero Gaussian random field with a covariance function

$$\gamma_0((s_1, s_2), (t_1, t_2)) = \Gamma_0(s_1, s_2, t_1, t_2) = \sum_{k=1}^{R} k^{-2} \psi_k(s_1, s_2) \psi_k(t_1, t_2),$$

where the eigenfunctions $\psi_k(t_1, t_2) \in \mathcal{P}_{r_1, r_2} := \{e_i(t_1) e_j(t_2) : i = 1, \ldots, r_1; j = 1, \ldots, r_2\}$. Three combinations of one-way ranks (r_1, r_2) and two-way rank R were studied for Γ_0:

$$\textbf{Setting 1}\text{: } R = 6,\ r_1 = 3,\ r_2 = 2; \qquad \textbf{Setting 2}\text{: } R = 6,\ r_1 = r_2 = 4;$$
$$\textbf{Setting 3}\text{: } R = r_1 = r_2 = 4.$$

For each setting, we chose R functions out of $\mathcal{P}_{r_1, r_2}$ to be $\{\psi_k\}$ such that smoother functions are associated with larger eigenvalues. The details are described in Section S2.1 of the supplementary material.

In terms of sampling plans, we considered both sparse and dense designs. Due to the space limit, here we only show and discuss the results for the sparse design, while defer those for the dense design to Section S2.3 of the supplementary material. For the sparse design, the random locations $\boldsymbol{T}_{ij}, j = 1, \ldots, m$, were independently generated from the continuous uniform distribution on $[0, 1]^2$ within each field and across different fields, and the random errors $\{\epsilon_{ij} : i = 1, \ldots, n; j = 1, \ldots, m\}$ were independently generated from $N(0, \sigma^2)$. In each of the 200 simulation runs, the observed data were obtained following (1) with various combinations of $m = 10, 20$, $n = 100, 200$ and noise level $\sigma = 0.1, 0.4$.

We compared the proposed method, denoted by mOpCov, with three existing methods: 1) OpCov: the estimator based on Wong and Zhang (2019) with adaption to two dimensional case (see Section 2); 2) ll-smooth: local linear smoothing with Epanechnikov kernel; 3) ll-smooth+: the two-step estimator constructed by retaining eigen-components of ll-smoothselected by 99% fraction of variation explained (FVE), and then removing the eigen-components with negative eigenvalues.

For both OpCov and mOpCov, 5-fold cross-validation was adopted to select the corresponding tuning parameters.

Table 1 show the average integrated squared error (AISE), average of estimated two-way rank ($\bar{R}$), as well as average of estimated one-way ranks ($\bar{r}_1, \bar{r}_2$) of the above covariance estimators over 200 simulated data sets in respective settings when sample size is $n = 200$. Corresponding results for $n = 100$ can be found in Table S4 of the supplementary material, and they lead to similar conclusions. Obviously ll-smooth and ll-smooth+, especially ll-smooth, perform significantly worse than the other two methods in both estimation accuracy and rank reduction (if applicable). Below we only compare mOpCov and OpCov.

Regarding estimation accuracy, the proposed mOpCov has uniformly smaller AISE values than OpCov, with around $10\% \sim 20\%$ improvement of AISE over OpCov in most cases under Settings 1 and 2. If the standard error (SE) of AISE is taken into account, the improvements of AISE by mOpCov are more distinguishable in Settings 1 and 2 than those in Setting 3 since the SEs of OpCov in Setting 3 are relatively high. This is due to the fact that in Setting 3, marginal basis are not shared by different two-dimensional eigenfunctions, and hence mOpCov cannot benefit from the structure sharing among eigenfunctions. Setting 3 is in fact an extreme setting we designed to challenge the proposed method.

For rank estimation, OpCov almost always underestimates two-way ranks, while mOpCov typically overestimates both one-way and two-way ranks. For mOpCov, the average one-way rank estimates are always smaller than the average two-way rank estimates, and their differences are substantial in Settings 1 and 2. This demonstrates the benefit of mOpCov of detecting structure sharing of one-dimensional basis among two-dimensional eigenfunctions.

We also tested the performance of mOpCov in the dense and regular designs, and compared it with the existing methods mentioned above together with the one by Wang and Huang (2017), which is not applicable to the sparse design. Details are given in Section S2.3 of the supplementary material, where all methods achieve similar AISE values, but mOpCov performs slightly better in estimation accuracy when the noise level is high.

7 Real Data Application

We applied the proposed method to an Argo profile data set, obtained from http://www.argo.ucsd.edu/Argo_data_and.html. The Argo project has a global array of approximately 3,800 free-drifting profiling floats, which measure temperature and salinity of the ocean. These floats drift freely in the depths of the ocean most of the time, and ascend regularly to the sea surface, where they transmit the collected data to the satellites. Every day only a small subset of floats show up on the sea surface. Due to the drifting process, these floats measure temperature and salinity at irregular locations over the ocean. See Figure 2 for examples.

In this analysis, we focus on the different changes of sea surface temperature between the tropical western and eastern Indian Ocean, which is called the Indian Ocean Dipole (IOD). The IOD is known to be associated with droughts in Australia (Ummenhofer et al., 2009) and has a significant effect on rainfall patterns in southeast Australia (Behera and Yamagata, 2003). According to Shinoda et al. (2004), the IOD phenomenon is a predominant inter-annual variation of sea surface temperature during late boreal summer and autumn (Shinoda et al., 2004), so in this application we focused on the sea surface temperature in the Indian Ocean region of longitude 40∼120 and latitude -20∼20 between September and November every year from 2003 to 2018.

Based on a simple autocorrelation analysis on the gridded data, we decided to use measurements

for every ten days in order to reduce the temporal dependence among the data.

At each location of a float on a particular day, the average temperature between 0 and 5 hPa from the float is regarded as a measurement. The Argo float dataset provides multiple versions of data, and we adopted the quality controlled (QC) version. Eventually we have a two-dimensional functional data collected of $n = 144$ days, where the number of observed locations $\boldsymbol{T}_{ij} = (\text{longitudie}, \text{latitude})$ per day varies from 7 to 47, i.e., $7 \leq m_i \leq 47$, $i = 1, ..., n$, with an average of 21.83. The locations are rescaled to $[0, 1] \times [0, 1]$. As shown in Figure 2, the data has a random sparse design.

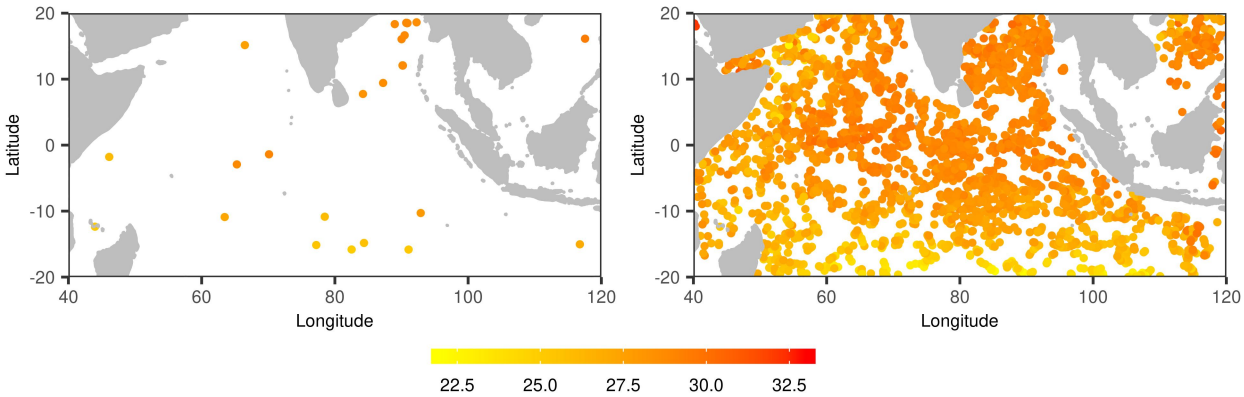


Figure 2: Observations on 2013/09/04 (left), and all observations in the data set (right). Points on the map indicate locations (Longitude, Latitude) of observations and the color scale of every point shows the corresponding Celsius temperature.

First we used kernel ridge regression with the corresponding kernel for the tensor product of two second order Sobolev spaces (e.g., Wong and Zhang, 2019) to obtain a mean function estimate for every month. Then we applied the proposed covariance function estimator with the same kernel.

The estimates of the top two two-dimensional L^2 eigenfunctions are illustrated in Figure 3. The first eigenfunction shows the east-west dipole mode, which aligns with existing scientific findings (e.g., Shinoda et al., 2004; Chu et al., 2014; Deser et al., 2010). The second eigenfunction can be interpreted as the basin-wide mode, which is a dominant mode all around the year (e.g., Deser et al., 2010; Chu et al., 2014).

To provide a clearer understanding of the covariance function structure, we derived a marginal L^2 basis along longitude and latitude respectively. The details are given in Appendix A. The left panel of Figure 4 demonstrates that the first longitudinal marginal basis reflects a large variation in the western region while the second one corresponds to the variation in the eastern region. Due to different linear combinations, the variation along longitude may contribute to not only opposite changes between the eastern and western sides of the Indian Ocean as shown in the first two-dimensional eigenfunction, but also an overall warming or cooling tendency as shown in the second two-dimensional eigenfunction. The second longitudinal marginal basis reveals that the closer to the east boundary, the greater the variation is, which suggests that the IOD may be related to the Pacific Ocean. This aligns with the evidence that the IOD has a link with El Niño Southern Oscillation (ENSO) (Stuecker et al., 2017), an irregularly periodic variation in sea surface temperature over the tropical eastern Pacific Ocean. As shown in the right panel of Figure 4, the overall trend for the first latitude marginal basis is almost a constant function. This provides

evidence that the IOD is primarily associated with the variation along longitude.

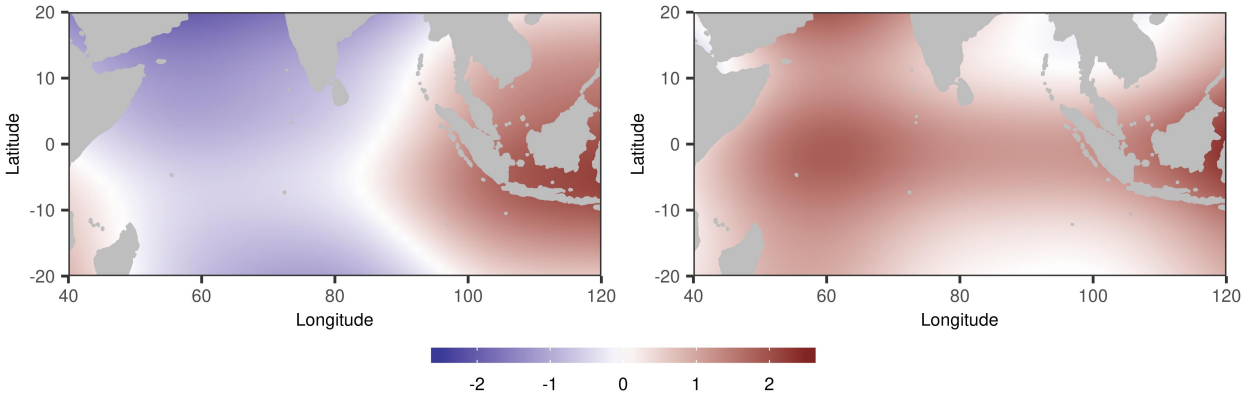
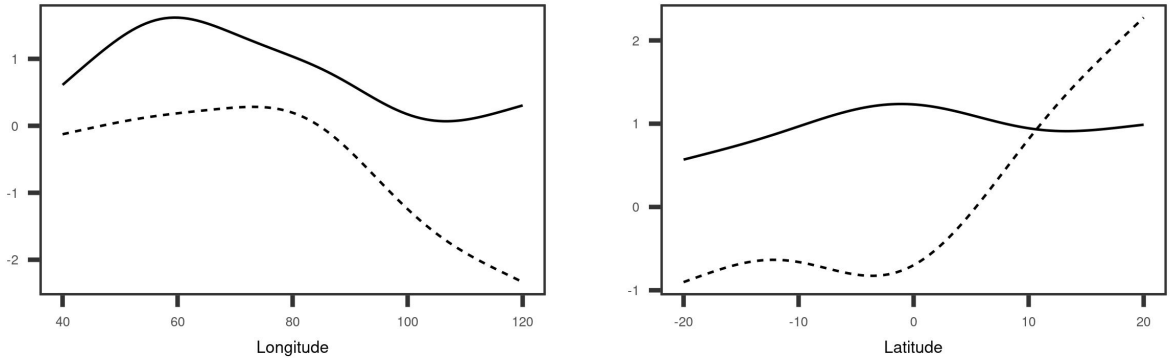


Figure 3: The first two-dimensional L^2 eigenfunction (left) and the second two-dimensional L^2 eigenfunction (right). The first eigenfunction explains 33.60% variance and the second eigenfunction explains 25.94% variance.



(a) Longitude (39.06%, 36.10%)

(b) Latitude (48.22%, 25.40%)

Figure 4: The first two marginal L^2 basis functions along longitude and latitude respectively. Solid lines are the first marginal basis function and dotted lines are the second marginal basis function. The fractions of variation explained by the corresponding principle components are given in parentheses.

Algorithm 1: Accelerated ADMM for solving (11)

Input: $\hat{\boldsymbol{V}}_k^{(0)} \in \mathbb{R}^{q_1 \times \cdots \times q_{2p}}$, $k = 0, 1, \ldots, p$, and $\boldsymbol{B}^{(0)} \in \mathbb{R}^{q_1 \times \cdots \times q_{2p}}$ such that $\hat{\boldsymbol{V}}_{0,(0)}$ and $\boldsymbol{B}_{\blacksquare}^{(0)}$ are symmetric matrices; $\boldsymbol{M}_k = [\boldsymbol{M}_{1,k}^\intercal, \ldots, \boldsymbol{M}_{n,k}^\intercal]^\intercal$, $k = 1, \ldots, p$; $\boldsymbol{Z}_i = (Z_{ijj'})_{1 \le j,j' \le m}$, $i = 1, \ldots, n$; $\tilde{\boldsymbol{I}} = [I(i \neq j)]_{1 \le i,j \le m}$; $\eta > 0$; T

Initialization: $\alpha_k^{(0)} \leftarrow 1$, $\boldsymbol{D}_k^{(-1)} \leftarrow \boldsymbol{B}^{(0)}$, $\hat{\boldsymbol{D}}_k^{(0)} \leftarrow \boldsymbol{B}^{(0)}$, $\boldsymbol{V}_k^{(-1)} \leftarrow \hat{\boldsymbol{V}}_k^{(0)}$, $k = 0, 1, \ldots, p$
$\boldsymbol{L}_i \leftarrow [\boldsymbol{M}_{i,1}^\intercal \odot \boldsymbol{M}_{i,2}^\intercal \odot \cdots \odot \boldsymbol{M}_{i,p}^\intercal]^\intercal$, $i = 1, \ldots, n$, where $\odot$ is the KhatriRao product defined as $\boldsymbol{A} \odot \boldsymbol{B} = [a_i \otimes b_i]_{i=1,\ldots,r} \in \mathbb{R}^{r_a r_b \times r}$ for $\boldsymbol{A} \in \mathbb{R}^{r_a \times r}$, $\boldsymbol{B} \in \mathbb{R}^{r_b \times r}$ and a_i, b_i are i-th column of matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ respectively.
$\boldsymbol{G} \leftarrow \frac{1}{nm(m-1)} \sum_{i=1}^n (\boldsymbol{L}_i \otimes \boldsymbol{L}_i)^\intercal \mathrm{diag}(\mathrm{vec}(\tilde{\boldsymbol{I}}))(\boldsymbol{L}_i \otimes \boldsymbol{L}_i)$
$\boldsymbol{h} \leftarrow \frac{2}{nm(m-1)} \sum_{i=1}^n (\boldsymbol{L}_i \otimes \boldsymbol{L}_i)^\intercal \mathrm{diag}(\mathrm{vec}(\tilde{\boldsymbol{I}}))\mathrm{vec}(\boldsymbol{Z}_i)$
$\boldsymbol{Q} \leftarrow (2(\boldsymbol{G} + \frac{p+1}{2} * \eta * \boldsymbol{I}))^{-1}$

for $t = 0, 1, \ldots, T$ **do**
  $\mathrm{vec}(\boldsymbol{B}_{\blacksquare}^{(t+1)}) \leftarrow \boldsymbol{Q}\{\boldsymbol{h} + \eta \sum_{k=0}^p \mathrm{vec}([\boldsymbol{D}_k^{(t)} - \hat{\boldsymbol{V}}_k^{(t)}]_{\blacksquare})\}$
  for $k = 0, 1, \ldots, p$ **do**
    if $k = 0$ **then**
      $\boldsymbol{D}_0^{(t)} \leftarrow \mathrm{prox}^+_{\lambda\beta/\eta}(\boldsymbol{B}^{(t+1)} + \hat{\boldsymbol{V}}_0^{(t)})$
    else
      $\boldsymbol{D}_k^{(t)} \leftarrow \mathrm{prox}^k_{\lambda(1-\beta)/(p\eta)}(\boldsymbol{B}^{(t+1)} + \hat{\boldsymbol{V}}_k^{(t)})$
    end
    $\boldsymbol{V}_k^{(t)} \leftarrow \hat{\boldsymbol{V}}_k^{(t)} + \boldsymbol{B}^{(t+1)} - \boldsymbol{D}_k^{(t)}$
    $\alpha_k^{(t+1)} \leftarrow \frac{1+\sqrt{1+4(\alpha_k^{(t)})^2}}{2}$
    $\hat{\boldsymbol{D}}_k^{(t+1)} \leftarrow \boldsymbol{D}_k^{(t)} + \frac{\alpha_k^{(t)}-1}{\alpha_k^{(t+1)}}(\boldsymbol{D}_k^{(k)} - \boldsymbol{D}_k^{(k-1)})$
    $\hat{\boldsymbol{V}}_k^{(t+1)} \leftarrow \boldsymbol{V}_k^{(t)} + \frac{\alpha_k^{(t)}-1}{\alpha_k^{(t+1)}}(\boldsymbol{V}_k^{(t)} - \boldsymbol{V}_k^{(t-1)})$
  end
  Stop if objective value change less than tolerance.
end

Output: $\boldsymbol{D}_0^{(T)}$

Table 1: Simulation results for three Settings with the sparse design when sample size (n) is 200. The AISE values with standard errors (SE) in parentheses are provided for the four covariance estimators in comparison, together with average two-way ranks ($\bar{R}$) for those estimators which can lead to rank reduction (i.e., mOpCov, OpCov, and ll-smooth+) and average one-way ranks (r_1, r_2) for mOpCov.

Setting	m	σ		mOpCov	OpCov	ll-smooth	ll-smooth+
1	10	0.1	AISE	0.053 (1.97e-03)	0.0632 (3.22e-03)	0.652 (1.92e-01)	0.337 (5.35e-02)
			$\bar{R}$	8.38	2.94	-	172.70
			$\bar{r}_1$, $\bar{r}_2$	5.4, 5.4	_	_	_
		0.4	AISE	0.0547 (2.01e-03)	0.0656 (2.72e-03)	0.714 (2.11e-01)	0.366 (5.96e-02)
			$\bar{R}$	9.16	2.84	-	177.3
			$\bar{r}_1$, $\bar{r}_2$	5.34, 5.32	_	_	_
	20	0.1	AISE	0.0343 (1.46e-03)	0.0421 (1.97e-03)	0.297 (1.39e-02)	0.206 (4.62e-03)
			$\bar{R}$	8.38	3.78	-	317.44
			$\bar{r}_1$, $\bar{r}_2$	5.84, 5.82	_	_	_
		0.4	AISE	0.0354 (1.52e-03)	0.044 (2.21e-03)	0.325 (1.58e-02)	0.223 (4.94e-03)
			$\bar{R}$	8.86	3.76	-	326.31
			$\bar{r}_1$, $\bar{r}_2$	5.83, 5.84	_	_	_
2	10	0.1	AISE	0.0532 (1.98e-03)	0.0636 (3.12e-03)	2.33 (1.13e+00)	0.795 (2.98e-01)
			$\bar{R}$	8.48	3.02	-	191.175
			$\bar{r}_1$, $\bar{r}_2$	5.82, 5.82	_	_	_
		0.4	AISE	0.0548 (2.05e-03)	0.0686 (3.53e-03)	2.44 (1.17e+00)	0.828 (3.04e-01)
			$\bar{R}$	9.04	3.04	-	196.34
			$\bar{r}_1$, $\bar{r}_2$	5.71, 5.74	_	_	_
	20	0.1	AISE	0.0341 (1.43e-03)	0.0419 (2.02e-03)	0.301 (1.58e-02)	0.208 (4.50e-03)
			$\bar{R}$	8.99	3.74	-	318.645
			$\bar{r}_1$, $\bar{r}_2$	5.93, 5.92	_	_	_
		0.4	AISE	0.0348 (1.43e-03)	0.043 (2.22e-03)	0.328 (1.78e-02)	0.225 (4.74e-03)
			$\bar{R}$	8.01	3.6	-	327.395
			$\bar{r}_1$, $\bar{r}_2$	5.94, 5.93	_	_	_
3	10	0.1	AISE	0.058 (2.62e-03)	0.0692 (5.33e-03)	0.454 (7.28e-02)	0.286 (2.89e-02)
			$\bar{R}$	6.26	3.12	-	182.74
			$\bar{r}_1$, $\bar{r}_2$	5, 5.06	_	_	_
		0.4	AISE	0.0598 (2.68e-03)	0.0733 (6.14e-03)	0.531 (1.07e-01)	0.323 (4.23e-02)
			$\bar{R}$	6.48	3.2	-	185.82
			$\bar{r}_1$, $\bar{r}_2$	4.99, 5.07	_	_	_
	20	0.1	AISE	0.0422 (1.37e-03)	0.0535 (2.64e-03)	0.267 (5.04e-03)	0.196 (3.59e-03)
			$\bar{R}$	6.29	4.49	-	332.09
			$\bar{r}_1$, $\bar{r}_2$	5.62, 5.69	_	_	_
		0.4	AISE	0.0424 (1.30e-03)	0.0494 (2.42e-03)	0.292 (5.30e-03)	0.212 (3.72e-03)
			$\bar{R}$	5.68	3.36	-	338.725
			$\bar{r}_1$, $\bar{r}_2$	5.59, 5.66	_	_	_

Appendix

A L^2 eigensystem and L^2 marginal basis

In this section, we present a transformation procedure to produce L^2 eigenfunctions and corresponding eigenvalues from our estimator $\hat{\boldsymbol{B}}$ obtained by (11).

Let $\boldsymbol{Q}_k = [\int_{[0,1]} K(s, T_{ijk})K(s, T_{i'j'k})ds]_{1\le i,i'\le n, 1\le j,j'\le m}$, $k = 1, \ldots, p$. Then $\boldsymbol{Q}_k = \boldsymbol{M}_k \boldsymbol{R}_k \boldsymbol{M}_k^\intercal$, where $\boldsymbol{R}_k = [\int_{[0,1]} v_l(s)v_h(s)ds]_{1\le l,h\le q_k}$ and $\{v_l : l = 1, \ldots, q_k\}$ form a basis of $\mathcal{H}_k$, so $\boldsymbol{R}_k = \boldsymbol{M}_k^+ \boldsymbol{Q}_k (\boldsymbol{M}_k^+)^\intercal$. The L^2 eigenvalues of $\hat{\Gamma}_\blacksquare$ coincide with the eigenvalues of matrix $\hat{\boldsymbol{B}}^L_{\text{square}} := (\boldsymbol{R}_1 \otimes \ldots \otimes \boldsymbol{R}_p)^{1/2} \hat{\boldsymbol{B}}_\blacksquare [(\boldsymbol{R}_1 \otimes \ldots \otimes \boldsymbol{R}_p)^{1/2}]^\intercal$, and the number of nonzero eigenvalues is the same as the rank of $\hat{\boldsymbol{B}}_\blacksquare$. The L^2 eigenfunction $\hat{\phi}_l$ that corresponds to the l-th eigenvalue of $\hat{\Gamma}_\blacksquare$ can be expressed as $\hat{\phi}_l(s_1, ..., s_p) = \boldsymbol{u}_l^\intercal [\boldsymbol{z}_1(s_1) \otimes \ldots \otimes \boldsymbol{z}_p(s_p)]$, where $\boldsymbol{z}_k(\cdot)$, $k = 1, \ldots, p$ are defined in Theorem 1, and $\boldsymbol{u}_l = (\boldsymbol{M}_1^+ \otimes \ldots \otimes \boldsymbol{M}_p^+)^\intercal (\boldsymbol{R}_1 \otimes \ldots \otimes \boldsymbol{R}_p)^{-1/2} \boldsymbol{v}_l$ with $\boldsymbol{v}_l$ being the l-th eigenvector of matrix $\hat{\boldsymbol{B}}^L_{\text{square}}$. Using the property of Kronecker products, we have $\hat{\phi}_l(s_1, ..., s_p) = \boldsymbol{v}_l^\intercal [(\boldsymbol{R}_1^{-1/2} \boldsymbol{M}_1^+ \boldsymbol{z}_1(s_1)) \otimes \ldots \otimes (\boldsymbol{R}_p^{-1/2} \boldsymbol{M}_p^+ \boldsymbol{z}_p(s_p))]$.

By simple verification, we can see that $\boldsymbol{R}_k^{-1/2} \boldsymbol{M}_k^+ \boldsymbol{z}_k(\cdot)$ are q_k one-dimensional orthonormal L^2 functions for dimension k, $k = 1, ..., p$. Therefore, we can also express $\hat{\Gamma}$ with these L^2 one-dimensional basis and the coefficients will form a $2p-$th order tensor of dimension $q_1 \times \ldots q_p \times q_1 \times \ldots q_p$. We use $\hat{\boldsymbol{B}}^L$ to represent this new coefficient tensor and extend our unfolding operators to L^2 space. It is easy to see that $\hat{\boldsymbol{B}}^L_\blacksquare = \hat{\boldsymbol{B}}^L_{\text{square}}$.

Since $\hat{\Gamma}_{(k)}$ is a compact operator in the L^2 space, this yields a singular value decomposition which leads to a L^2 basis characterizing the marginal variation along the $k-$th dimension. We call it a L^2 *marginal basis* for the $k-$th dimension. Obviously the marginal basis function $\hat{\psi}_l^k$ corresponding to the l-th singular value for dimension k can be expressed as $\hat{\psi}_l^k(\cdot) = \boldsymbol{u}_l^k \boldsymbol{z}_k(\cdot)$, where $\boldsymbol{u}_l^k = (\boldsymbol{M}_k^+)^\intercal \boldsymbol{R}_k^{-1/2} \boldsymbol{v}_l^k$, and $\boldsymbol{v}_l^k$ is the l-th singular vector of $\hat{\boldsymbol{B}}^L_{(k)}$. And the L^2 marginal singular values of $\hat{\Gamma}_{(k)}$ coincide with the singular values of matrix $\hat{\boldsymbol{B}}^L_{(k)}$.

B Definitions of $\kappa_{n,m}$ and $\eta_{n,m}$

Here we provide the specific forms of $\kappa_{n,m}$ and $\eta_{n,m}$, which are closely related to the decay of $\{\mu_l \mu_h : l, h = 1, \ldots\}$. Specifically, $\kappa_{n,m}$ is defined as the smallest positive κ such that

$$\begin{aligned} cb^3 \left[\frac{1}{n(m-1)} \sum_{l,h=1}^{\infty} \min\left\{\kappa^2, \mu_l \mu_h\right\} \right]^{1/2} &\le \kappa^2, \\ 32cb \left[\frac{1}{n(m-1)} \sum_{l,h=1}^{\infty} \min\left\{\kappa^2/b^2, \mu_l \mu_h\right\} \right]^{1/2} &\le \kappa^2, \end{aligned} \tag{17}$$

where c is a universal constant, and $\eta_{n,m}$ is defined as the smallest positive η such that

$$\left(\frac{c_\eta}{nm} \sum_{l,h=1}^{\infty} \min\{\eta^2, \mu_l \mu_h\} + \frac{\eta^2}{n} \right)^{1/2} \le \eta^2, \tag{18}$$

where c_η is a constant depending on b, b_X, b_ϵ. The existences of $\kappa_{n,m}$ and $\eta_{n,m}$ are shown in the proof of Theorem 2.

Supplementary Material

In the supplementary material related to this paper, we provide proofs of our theoretical findings and additional simulation results.

Acknowledgement

The research of Raymond K. W. Wong is partially supported by the U.S. National Science Foundation under grants DMS-1806063, DMS-1711952 (subcontract) and CCF-1934904. The research of Xiaoke Zhang is partially supported by the U.S. National Science Foundation under grant DMS-1832046. Portions of this research were conducted with high performance research computing resources provided by Texas A&M University (https://hprc.tamu.edu).

References

Abernethy, J., F. Bach, T. Evgeniou, and J.-P. Vert (2009). A new approach to collaborative filtering: Operator estimation with spectral regularization. *Journal of Machine Learning Research 10*, 803–826.

Allen, G. I. (2013). Multi-way functional principal components analysis. In *2013 5th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pp. 220–223. IEEE.

Bartlett, P. L., O. Bousquet, and S. Mendelson (2005). Local rademacher complexities. *The Annals of Statistics 33*(4), 1497–1537.

Behera, S. K. and T. Yamagata (2003). Influence of the indian ocean dipole on the southern oscillation. *Journal of the Meteorological Society of Japan. Ser. II 81*(1), 169–177.

Boyd, S., N. Parikh, E. Chu, B. Peleato, and J. Eckstein (2010). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning 3*(1), 1–122.

Cai, T. T. and M. Yuan (2010). Nonparametric covariance function estimation for functional and longitudinal data. Technical report, Georgia Institute of Technology, Atlanta, GA.

Cai, T. T. and M. Yuan (2012). Minimax and adaptive prediction for functional linear regression. *Journal of the American Statistical Association 107*(499), 1201–1216.

Chen, K., P. Delicado, and H.-G. Müller (2017). Modelling function-valued stochastic processes, with applications to fertility dynamics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79*(1), 177–196.

Chen, K. and H.-G. Müller (2012). Modeling repeated functional observations. *Journal of the American Statistical Association 107*(500), 1599–1609.

Chen, L.-H. and C.-R. Jiang (2017). Multi-dimensional functional principal component analysis. *Statistics and Computing 27*(5), 1181–1192.

Chu, J.-E., K.-J. Ha, J.-Y. Lee, B. Wang, B.-H. Kim, and C. E. Chung (2014). Future change of the indian ocean basin-wide and dipole modes in the cmip5. *Climate dynamics 43*(1-2), 535–551.

Deser, C., M. A. Alexander, S.-P. Xie, and A. S. Phillips (2010). Sea surface temperature variability: Patterns and mechanisms. *Annual review of marine science 2*, 115–143.

Ferraty, F. and P. Vieu (2006). *Nonparametric functional data analysis: theory and practice*. Springer, New York.

Goldsmith, J., J. Bobb, C. M. Crainiceanu, B. Caffo, and D. Reich (2011). Penalized functional regression. *Journal of Computational and Graphical Statistics 20*(4), 830–851.

Gu, C. (2013). *Smoothing Spline ANOVA Models* (2nd ed.). New York: Springer.

Hackbusch, W. (2012). *Tensor spaces and numerical tensor calculus*, Volume 42. Springer Science & Business Media.

Halko, N., P.-G. Martinsson, and J. A. Tropp (2009). Finding structure with randomness: Stochastic algorithms for constructing approximate matrix decompositions. *arXiv preprint arXiv:0909.4061*.

Hall, P. and C. Vial (2006). Assessing the finite dimensionality of functional data. *Journal of the Royal Statistical Society: Series B 68*(4), 689–705.

Horváth, L. and P. Kokoszka (2012). *Inference for functional data with applications*, Volume 200. Springer, New York.

Hsing, T. and R. Eubank (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operators*. John Wiley & Sons.

Huang, J. Z., H. Shen, and A. Buja (2009). The analysis of two-way functional data using two-way regularized singular value decompositions. *Journal of the American Statistical Association 104*(488), 1609–1620.

James, G. M., T. J. Hastie, and C. A. Sugar (2000). Principal component models for sparse functional data. *Biometrika 87*(3), 587–602.

Kadkhodaie, M., K. Christakopoulou, M. Sanjabi, and A. Banerjee (2015). Accelerated alternating direction method of multipliers. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 497–506. ACM.

Kokoszka, P. and M. Reimherr (2017). *Introduction to functional data analysis*. CRC Press.

Koltchinskii, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*, Volume 2033. Springer Science & Business Media.

Li, B. and J. Song (2017). Nonlinear sufficient dimension reduction for functional data. *The Annals of Statistics 45*(3), 1059–1095.

Li, Y. and T. Hsing (2010). Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *The Annals of Statistics 38*(6), 3321–3351.

Liebl, D. (2019). Inference for sparse and dense functional data with covariate adjustments. *Journal of Multivariate Analysis 170*, 315–335.

Lynch, B. and K. Chen (2018). A test of weak separability for multi-way functional data, with application to brain connectivity studies. *Biometrika 105*(4), 815–831.

Mazumder, R., T. Hastie, and R. Tibshirani (2010). Spectral regularization algorithms for learning large incomplete matrices. *Journal of machine learning research 11*(Aug), 2287–2322.

Mercer, J. (1909). Xvi. functions of positive and negative type, and their connection the theory of integral equations. *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character 209*(441-458), 415–446.

Park, S. Y. and A.-M. Staicu (2015). Longitudinal functional data analysis. *Stat 4*(1), 212–226.

Paul, D. and J. Peng (2009). Consistency of restricted maximum likelihood estimators of principal components. *The Annals of Statistics 37*(3), 1229–1271.

Pearce, N. D. and M. P. Wand (2006). Penalized splines and reproducing kernel methods. *The American Statistician 60*(3), 233–240.

Poskitt, D. S. and A. Sengarapillai (2013). Description length and dimensionality reduction in functional data analysis. *Computational Statistics & Data Analysis 58*, 98–113.

Ramsay, J. and B. Silverman (2005). *Functional data analysis*. Springer, New York.

Raskutti, G., M. J. Wainwright, and B. Yu (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *The Journal of Machine Learning Research 13*, 389–427.

Reimherr, M., B. Sriperumbudur, and B. Taoufik (2018). Optimal prediction for additive function-on-function regression. *Electronic Journal of Statistics 12*(2), 4571–4601.

Rice, J. A. and B. W. Silverman (1991). Estimating the mean and covariance structure non-parametrically when the data are curves. *Journal of the Royal Statistical Society: Series B (Methodological) 53*(1), 233–243.

Shamshoian, J., D. Senturk, S. Jeste, and D. Telesca (2019). Bayesian analysis of multidimensional functional data. *arXiv preprint arXiv:1909.08763*.

Shinoda, T., H. H. Hendon, and M. A. Alexander (2004). Surface and subsurface dipole variability in the indian ocean and its relation with enso. *Deep Sea Research Part I: Oceanographic Research Papers 51*(5), 619–635.

Stuecker, M. F., A. Timmermann, F.-F. Jin, Y. Chikamoto, W. Zhang, A. T. Wittenberg, E. Widiasih, and S. Zhao (2017). Revisiting enso/indian ocean dipole phase relationships. *Geophysical Research Letters 44*(5), 2481–2492.

Sun, X., P. Du, X. Wang, and P. Ma (2018). Optimal penalized function-on-function regression under a reproducing kernel hilbert space framework. *Journal of the American Statistical Association 113*(524), 1601–1611.

Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika 31*(3), 279–311.

Ummenhofer, C. C., M. H. England, P. C. McIntosh, G. A. Meyers, M. J. Pook, J. S. Risbey, A. S. Gupta, and A. S. Taschetto (2009). What causes southeast australia's worst droughts? *Geophysical Research Letters 36*(4).

Wahba, G. (1990). *Spline Models for Observational Data*. Philadelphia: SIAM.

Wang, W.-T. and H.-C. Huang (2017). Regularized principal component analysis for spatial data. *Journal of Computational and Graphical Statistics 26*(1), 14–25.

Wong, R. K. W., Y. Li, and Z. Zhu (2019). Partially linear functional additive models for multivariate functional data. *Journal of the American Statistical Association 114*(525), 406–418.

Wong, R. K. W. and X. Zhang (2019). Nonparametric operator-regularized covariance function estimation for functional data. *Computational Statistics & Data Analysis 131*, 131–144.

Xiao, L., Y. Li, and D. Ruppert (2013). Fast bivariate p-splines: the sandwich smoother. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75*(3), 577–599.

Yao, F., H.-G. Müller, and J.-L. Wang (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association 100*(470), 577–590.

Yuan, M. and T. T. Cai (2010). A reproducing kernel hilbert space approach to functional linear regression. *The Annals of Statistics 38*(6), 3412–3444.

Zhang, L., H. Shen, and J. Z. Huang (2013). Robust regularized singular value decomposition with application to mortality data. *The Annals of Applied Statistics 7*(3), 1540–1561.

Zhang, X. and J.-L. Wang (2016). From sparse to dense functional data and beyond. *The Annals of Statistics 44*(5), 2281–2321.

Zhou, L. and H. Pan (2014). Principal component analysis of two-dimensional functional data. *Journal of Computational and Graphical Statistics 23*(3), 779–801.

Zhu, H., F. Yao, and H. H. Zhang (2014). Structured functional additive regression in reproducing kernel hilbert spaces. *Journal of the Royal Statistical Society: Series B 76*(3), 581–603.

Zipunnikov, V., B. Caffo, D. M. Yousem, C. Davatzikos, B. S. Schwartz, and C. Crainiceanu (2011). Multilevel functional principal component analysis for high-dimensional data. *Journal of Computational and Graphical Statistics 20*(4), 852–873.

Supplementary material for "Low-Rank Covariance Function Estimation for Multidimensional Functional Data"

Jiayi Wang[1], Raymond K. W. Wong*[1], and Xiaoke Zhang†[2]

[1]*Department of Statistics, Texas A&M University*
[2]*Department of Statistics, George Washington University*

September 1, 2020

S1 Proofs

S1.1 Proof of Theorem 1

For any $\Gamma \in \mathcal{G}$, we can decompose it into two orthogonal parts Γ_1 and Γ_2 such that $\Gamma_1 \in \mathcal{G}(\mathcal{L}_{n,m})$ and $\Gamma_2 \in (\mathcal{G}(\mathcal{L}_{n,m}))^{\perp}$. Since the loss function $\ell(\Gamma)$ only depends on data, it suffices to show that $\Psi_0(\Gamma_{\blacksquare}) \geq \Psi_0(\Gamma_{1,\blacksquare})$ and $\Psi_k(\Gamma_{(k)}) \geq \Psi_k(\Gamma_{1,(k)})$ for $k = 1, \ldots, p$. Below we follow two steps to prove this.

Step 1. Take $\mathcal{H}(\mathcal{L}_{n,m}) := \bigotimes_{k=1}^{p} \mathcal{K}_k$. Since we require $\Gamma \in \mathcal{M}^+$, we first show that $\Gamma_{1,\blacksquare} = \Gamma_{1,\blacksquare}^{\intercal}$ and $\langle \Gamma_{1,\blacksquare} f, f \rangle_{\mathcal{H}} \geq 0$ for any $f \in \mathcal{H}$. Note that $\Gamma_{\blacksquare} = \Gamma_{\blacksquare}^{\intercal}$, so $\Gamma_{\blacksquare} = (\Gamma_{1,\blacksquare} + \Gamma_{2,\blacksquare})/2 + (\Gamma_{1,\blacksquare}^{\intercal} + \Gamma_{2,\blacksquare}^{\intercal})/2$. As $\Gamma_{1,\blacksquare}^{\intercal} \in \mathcal{H}(\mathcal{L}_{n,m}) \otimes \mathcal{H}(\mathcal{L}_{n,m})$ and $\Gamma_{2,\blacksquare}^{\intercal} \in (\mathcal{H}(\mathcal{L}_{n,m}) \otimes \mathcal{H}(\mathcal{L}_{n,m}))^{\perp}$, we have $\Gamma_1 = (\Gamma_{1,\blacksquare} + \Gamma_{1,\blacksquare}^{\intercal})/2$ and $\Gamma_2 = (\Gamma_{2,\blacksquare} + \Gamma_{2,\blacksquare}^{\intercal})/2$. Thus $\Gamma_{1,\blacksquare} = \Gamma_{1,\blacksquare}^{\intercal}$ and $\Gamma_{2,\blacksquare} = \Gamma_{2,\blacksquare}^{\intercal}$.

By the definition of Γ_2, $\langle \Gamma_{2,\blacksquare} g, g \rangle_{\mathcal{H}} = 0$ for any $g \in \mathcal{H}(\mathcal{L}_{n,m})$, so we have

$$0 \leq \langle \Gamma_{\blacksquare} g, g \rangle_{\mathcal{H}} = \langle \Gamma_{1,\blacksquare} g, g \rangle_{\mathcal{H}} + \langle \Gamma_{2,\blacksquare} g, g \rangle_{\mathcal{H}} = \langle \Gamma_{1,\blacksquare} g, g \rangle_{\mathcal{H}}.$$

Moreover, the definition of Γ_1 leads to $\langle \Gamma_{1,\blacksquare} g, g \rangle_{\mathcal{H}} = 0$ for any $g \in (\mathcal{H}(\mathcal{L}_{n,m}))^{\perp}$. Hence $\langle \Gamma_{1,\blacksquare} f, f \rangle_{\mathcal{H}} \geq 0$ for any $f \in \mathcal{H}$.

Step 2. Next we show that for all k, $\lambda_k(\Gamma_{\blacksquare}) \geq \lambda_k(\Gamma_{1,\blacksquare})$ and $\lambda_k(\Gamma_{(j)}) \geq \lambda_k(\Gamma_{1,(j)})$ with $j = 1, \ldots, p$. Let $P_{\mathcal{H}(\mathcal{L}_{n,m})}$ be the projection operator to space $\mathcal{H}(\mathcal{L}_{n,m})$ and denote the adjoint operator of A by A^*. Then we have

$$\begin{aligned} \lambda_k(\Gamma_{1,\blacksquare}) &= \lambda_k \left(P_{\mathcal{H}(\mathcal{L}_{n,m})} \Gamma_{\blacksquare} P_{\mathcal{H}(\mathcal{L}_{n,m})} \right) \\ &\leq \lambda_k \left(\Gamma_{\blacksquare} P_{\mathcal{H}(\mathcal{L}_{n,m})} \right) = \lambda_k \left(P_{\mathcal{H}(\mathcal{L}_{n,m})} \Gamma_{\blacksquare}^* \right) \leq \lambda_k \left(\Gamma_{\blacksquare}^* \right) = \lambda_k(\Gamma_{\blacksquare}). \end{aligned}$$

*The research of Raymond K. W. Wong is partially supported by National Science Foundation grants DMS-1806063, DMS-1711952 and CCF-1934904.

†The research of Xiaoke Zhang is partially supported by National Science Foundation grant DMS-1832046.

Let $P_{\mathcal{K}_j}$ denote the projection operator to space $\mathcal{K}_j$ and $P_{\mathcal{K}_{-j}}$ as the projection operator to space $\bigotimes_{k=1,k\neq j}^{2p}\mathcal{K}_k$ where $\mathcal{K}_{p+k}=\mathcal{K}_k$, $j=1,\ldots,p$. Then

$$\lambda_k(\Gamma_{1,(j)})=\lambda_k\left(P_{\mathcal{K}_j}\Gamma_{(j)}P_{\mathcal{K}_{-j}}\right)\leq\lambda_k\left(\Gamma_{(j)}P_{\mathcal{K}_{-j}}\right)=\lambda_k\left(P_{\mathcal{K}_{-j}}\Gamma_{(j)}^*\right)\leq\lambda_k\left(\Gamma_{(j)}^*\right)=\lambda_k\left(\Gamma_{(j)}\right).$$

Therefore, $\Psi_0(\Gamma_\blacksquare)\geq\Psi_0(\Gamma_{1,\blacksquare})$ and $\Psi_k(\Gamma_{(k)})\geq\Psi_k(\Gamma_{1,(k)})$ for $k=1,\ldots,p$.

S1.2 Proofs of Theorems 2, 3 and Corollary 1

For notational simplicity, we do not adopt different notations for the fully folded and squarely unfolded versions of operators (functions) in this section.

Write $\Delta=\hat{\Gamma}-\Gamma_0$ and $e(\boldsymbol{T}_{ij},\boldsymbol{T}_{ij'})=(X_i(\boldsymbol{T}_{ij})+\epsilon_{ij})(X_i(\boldsymbol{T}_{ij'})+\epsilon_{ij'})-\Gamma_0(\boldsymbol{T}_{ij},\boldsymbol{T}_{ij'})$. From (7), we obtain the following basic inequality:

$$\|\Delta\|_{n,m}^2+\lambda I(\hat{\Gamma})\leq 2\langle e,\Delta\rangle_{n,m}+\lambda I(\Gamma_0). \tag{S1}$$

The term $\langle e,\Delta\rangle_{n,m}$ involved in (S1) plays a crucial role in the subsequent asymptotic analysis, so we will focus on this term first.

Consider $\mathcal{G}_s=\{(\Gamma-\Gamma_0)/\{I(\Gamma)+I(\Gamma_0)\}:\Gamma\in\mathcal{G}\}$. To bound $\langle e,\Delta\rangle_{n,m}$, we start with controlling $\sup_{g\in\mathcal{G}_s}\langle e,g\rangle_{n,m}$. For any $g\in\mathcal{G}_s$, there exists a $\Gamma\in\mathcal{G}$ such that $g=(\Gamma-\Gamma_0)/\{I(\Gamma)+I(\Gamma_0)\}$. When $\Gamma=\Gamma_0$, $\|g\|_{\mathcal{G}}=0$. Otherwise,

$$\|g\|_{\mathcal{G}}=\left\|\frac{\Gamma-\Gamma_0}{I(\Gamma)+I(\Gamma_0)}\right\|_{\mathcal{G}}\leq\frac{\|\Gamma-\Gamma_0\|_{\mathcal{G}}}{I(\Gamma-\Gamma_0)}\leq\frac{\|\Gamma-\Gamma_0\|_{\mathcal{G}}}{\|\Gamma-\Gamma_0\|_{\mathcal{G}}}=1,$$

where the second inequality is due to that $I(\Gamma)\geq\|\Gamma\|_{\mathcal{G}}$ for any $\Gamma\in\mathcal{G}$, and $\|\cdot\|_{\mathcal{G}}$ is Hilbert–Schmidt norm of RKHS $\mathcal{G}$. Take $\mathcal{G}'=\{g\in\mathcal{G}:\|g\|_{\mathcal{G}}\leq 1\}$. From the above, one can easily see that $\mathcal{G}_s\subseteq\mathcal{G}'$, and hence $\sup_{g\in\mathcal{G}_s}\langle e,g\rangle_{n,m}\leq\sup_{g\in\mathcal{G}'}\langle e,g\rangle_{n,m}$ for any e. In the later part of our analysis, we will bound $\sup_{g\in\mathcal{G}'}\langle e,g\rangle_{n,m}$ to control $\sup_{g\in\mathcal{G}_s}\langle e,g\rangle_{n,m}$.

First, we note that the functions residing in $\mathcal{G}'$ are bounded: For any $g\in\mathcal{G}'$, by the property of reproducing kernel,

$$\sup_{g\in\mathcal{G}'}|g|_\infty\leq\sup_{(x_1,\ldots,x_{2p})\in[0,1]^{2p}}K((x_1,\ldots,x_{2p}),(x_1,\ldots,x_{2p}))\leq b.$$

Next we recall the definition of the sub-exponential norm of a random variable.

Definition S1. *For a random variable X, its sub-exponential norm is defined as*

$$\|X\|_{\psi_1}=\inf\{\lambda>0:\mathbb{E}(\exp(|X|/\lambda))\leq 2\}.$$

If $\|X\|_{\psi_1}<\infty$, then we call X a sub-exponential random variable.

Recall that $\mathcal{L}_{n,m}=\{T_{ijk}:i=1,\ldots,n;j=1,\ldots,m;k=1,\ldots,p\}$. We write $e_{ijj'}=e(\boldsymbol{T}_{ij},\boldsymbol{T}_{ij'})$. For random variables A and B, we denote by $\|A\mid B\|_{\psi_1}$ the sub-exponential norm of the random variable A conditional on B. The notation naturally extends to the case when B is a random vector or a set of random variables. By Lemma 3 in Wong and Zhang (2019), we can see that conditioned on $\mathcal{L}_{n,m}$, $e_{ijj'}$ are sub-exponential random variables. Moreover, there exists a constant σ_{ψ_1}, depending on b_X and b_ϵ, such that $\|e_{ijj'}\mid\mathcal{L}_{n,m}\|_{\psi_1}\leq\sigma_{\psi_1}^2$.

Next we introduce the following random variables:

$$\hat{Z}_{n,m}(e,t;\mathcal{G}') := \sup_{\{g\in\mathcal{G}':\|g\|_{n,m}\le t\}} \left| \frac{1}{nm(m-1)} \sum_{i=1}^{n}\sum_{j\neq j'}^{m} e_{ijj'} g(\boldsymbol{T}_{ij},\boldsymbol{T}_{ij'}) \right|,$$

$$\tilde{Z}_{n,m}(e,t;\mathcal{G}') := \sup_{\{g\in\mathcal{G}':\|g\|_{2}\le t\}} \left| \frac{1}{nm(m-1)} \sum_{i=1}^{n}\sum_{j\neq j'}^{m} e_{ijj'} g(\boldsymbol{T}_{ij},\boldsymbol{T}_{ij'}) \right|.$$

Our immediate goal is to bound $\hat{Z}_{n,m}(e,t;\mathcal{G}')$, which will be achieved by bounding $\tilde{Z}_{n,m}(e,t;\mathcal{G}')$. We start with its expectation. Without loss of generality, we use c to denote all the universal constants.

Lemma S1. *There exists a constant $c_\eta > 0$, depending on σ_{ψ_1} and L, such that*

$$\mathbb{E}\left[\left\{\tilde{Z}_{n,m}(e,t;\mathcal{G}')\right\}^2\right] \le c_\eta \left(\frac{1}{nm}\sum_{l,h=1}^{\infty}\min\{t^2,\mu_l\mu_h\} + \frac{t^2}{n}\right). \tag{S2}$$

Proof. A majority of the proof resembles that of Lemma 42 in Mendelson (2002), with additional arguments developed to control an important expectation term. Since the sample field of X resides in $\mathcal{H}$, we can decompose $X(\boldsymbol{t}) = \sum_{h=1}^{\infty}\zeta_h\phi_h(\boldsymbol{t})$ where $\mathbb{E}(\zeta_h\zeta_{h'}) = \mathbb{E}\left\{\Gamma_0(\boldsymbol{T}_{ij},\boldsymbol{T}_{ij'})\phi_h(\boldsymbol{T}_{ij})\phi_h'(\boldsymbol{T}_{ij'})\right\}$. For every $\boldsymbol{s},\boldsymbol{t}\in[0,1]^p$, write $\Phi(\boldsymbol{s},\boldsymbol{t}) = \left(\sqrt{\mu_l\mu_h}\phi_l(\boldsymbol{s})\phi_h(\boldsymbol{t})\right)_{l,h=1}^{\infty}$. For two squarely summable sequences $a=\{a_{lh}\}_{l,h=1}^{\infty}$ and $b=\{b_{lh}\}_{l,h=1}^{\infty}$, define their inner product and the 2-norm in the following: $\langle a,b\rangle = \sum_{l,h=1}^{\infty}a_{lh}b_{lh}$ and $\|a\|_2 = (\sum_{l,h=1}^{\infty}a_{lh}^2)^{1/2}$. One can show that

$$\mathcal{G}' = \{g(\cdot,\star) = \langle\beta,\,\Phi(\cdot,\star)\rangle : \|\beta\|_2\le 1\}.$$

Let $\mathcal{B}(t) = \{\beta : \|g\|_2\le t\}$. It follows that $g\in\mathcal{G}'\cap\mathcal{B}(t)$ if and only if β belongs to set $\Omega = \{\beta : \sum_{l,h=1}^{\infty}\beta_{lh}^2(\mu_l\mu_h)\le t^2, \sum_{l,h=1}^{\infty}\beta_{lh}^2\le 1\}$. Let $\Xi = \{\beta : \sum_{l,h=1}^{\infty}\beta_{lh}^2\nu_{lh}\le 1\}$, where $\nu_{lh} = (\min\{1,t^2/\mu_l\mu_h\})^{-1}$. We can see that $\Xi\subset\Omega\subset\sqrt{2}\Xi$, which implies

$$\mathbb{E}\left(\tilde{Z}_{n,m}(\omega,t;\mathcal{G}')\right)^2 \asymp \frac{1}{n^2m^2(m-1)^2}\mathbb{E}\sup_{\beta\in\Xi}\langle\beta,\sum_{i=1}^{n}\sum_{j\neq j'}^{m}e_{ijj'}\Phi(\boldsymbol{T}_{ij},\boldsymbol{T}_{ij'})\rangle^2.$$

Next,

$$\begin{aligned}
&\mathbb{E}\sup_{\beta\in\Xi}\left\langle\beta,\sum_{i=1}^{n}\sum_{j\neq j'}^{m}e_{ijj'}\Phi(\boldsymbol{T}_{ij},\boldsymbol{T}_{ij'})\right\rangle^2\\
&= \mathbb{E}\sup_{\beta\in\Xi}\left\langle\sum_{l,h=1}^{\infty}\sqrt{\nu_{lh}}\beta_{lh},\sum_{l,h=1}^{\infty}\frac{\sqrt{\mu_l\mu_h}}{\sqrt{\nu_{lh}}}\sum_{i=1}^{n}\sum_{j\neq j'}^{m}e_{ijj'}\phi_l(\boldsymbol{T}_{ij})\phi_h(\boldsymbol{T}_{ij'})\right\rangle^2\\
&\le \mathbb{E}\sum_{l,h=1}^{\infty}\frac{\mu_l\mu_h}{\nu_{lh}}\left\{\sum_{i=1}^{n}\sum_{j\neq j'}^{m}e_{ijj'}\phi_l(\boldsymbol{T}_{ij})\phi_h(\boldsymbol{T}_{ij'})\right\}^2\\
&= n\sum_{l,h=1}^{\infty}\frac{\mu_l\mu_h}{\nu_{lh}}\mathbb{E}\left\{\sum_{j\neq j'}^{m}e_{1jj'}\phi_l(\boldsymbol{T}_{1j})\phi_h(\boldsymbol{T}_{1j'})\right\}^2.
\end{aligned}$$

The last equality follows from the independence between different sample fields and observed locations, combined with the fact that $\mathbb{E}(e_{ijj'} \mid \mathcal{L}_{n,m}) = 0$.

It remains to bound $\mathbb{E}\left\{\sum_{j\neq j'}^{m} e_{1jj'}\phi_l(\boldsymbol{T}_{1j})\phi_h(\boldsymbol{T}_{1j'})\right\}^2$. Write

$$U_{jj'kk'} = e_{1jj'}e_{1kk'}\phi_l(\boldsymbol{T}_{1j})\phi_h(\boldsymbol{T}_{1j'})\phi_l(\boldsymbol{T}_{1k})\phi_h(\boldsymbol{T}_{1k'}).$$

When $j = k$ and $j' = k'$,

$$\begin{aligned} U_{jj'jj'} &= \mathbb{E}e_{1jj'}^2\phi_l^2(\boldsymbol{T}_{1j})\phi_h^2(\boldsymbol{T}_{1j'}) = \mathbb{E}\left[\left\{\mathbb{E}(e_{1jj'}^2 \mid \mathcal{L}_{n,m})\right\}\phi_l^2(\boldsymbol{T}_{1j})\phi_h^2(\boldsymbol{T}_{1j'})\right] \\ &\leq c\sigma_{\psi_1}^2\mathbb{E}\left\{\phi_l^2(\boldsymbol{T}_{1j})\phi_h^2(\boldsymbol{T}_{1j'})\right\} = c\sigma_{\psi_1}^2, \end{aligned}$$

where the inequality follows from the property of sub-exponential random variables and c is a universal constant. When $j = k$ and $j' \neq k'$,

$$\begin{aligned} U_{jj'jk'} &= \mathbb{E}\left\{e_{1jj'}e_{1jk'}\phi_l^2(\boldsymbol{T}_{1j})\phi_h(\boldsymbol{T}_{1j'})\phi_h(\boldsymbol{T}_{1k'})\right\} \\ &\leq \mathbb{E}\left[\left\{\mathbb{E}(e_{1jj'} \mid \mathcal{L}_{n,m})^2\mathbb{E}(e_{1jk'} \mid \mathcal{L}_{n,m})^2\right\}^{1/2}\phi_l^2(\boldsymbol{T}_{1j})\phi_h(\boldsymbol{T}_{1j'})\phi_h(\boldsymbol{T}_{1k'})\right] \\ &\leq c\sigma_{\psi_1}^2\mathbb{E}\left\{\phi_l^2(\boldsymbol{T}_{1j})\phi_h(\boldsymbol{T}_{1j'})\phi_h(\boldsymbol{T}_{1k'})\right\} \leq c\sigma_{\psi_1}^2\left\{\mathbb{E}\phi_h^2(\boldsymbol{T}_{1j'})\mathbb{E}\phi_h^2(\boldsymbol{T}_{1k'})\right\}^{1/2} \leq c\sigma_{\psi_1}^2. \end{aligned}$$

Similarly for $j \neq k$ and $j' = k'$, $U_{jj'kj'} \leq c\sigma_{\psi_1}^2$. When $j \neq k$ and $j' \neq k'$,

$$\begin{aligned} U_{jj'kk'} &= \mathbb{E}\left\{\mathbb{E}e_{1jj'}\phi_l(\boldsymbol{T}_{1j})\phi_h(\boldsymbol{T}_{1j'}) \mid X\right\}^2 \\ &= \mathbb{E}\left[\mathbb{E}\left\{(X(\boldsymbol{T}_{1j}) + \epsilon_{1j})(X(\boldsymbol{T}_{1j'}) + \epsilon_{1j'}) - \Gamma_0(\boldsymbol{T}_{1j}, \boldsymbol{T}_{1j'})\right\}\phi_l(\boldsymbol{T}_{1j})\phi_h(\boldsymbol{T}_{1j'}) \mid X\right]^2 \\ &= \mathbb{E}\left[\mathbb{E}\left\{X(\boldsymbol{T}_{1j})X(\boldsymbol{T}_{1j'})\phi_l(\boldsymbol{T}_{1j})\phi_h(\boldsymbol{T}_{1j'}) \mid X\right\} - \mathbb{E}\Gamma_0(\boldsymbol{T}_{1j}, \boldsymbol{T}_{1j'})\phi_l(\boldsymbol{T}_{1j})\phi_h(\boldsymbol{T}_{1j'})\right]^2 \\ &= \mathbb{E}\Bigg[\mathbb{E}\left\{\sum_{g=1}^{\infty}\zeta_g\phi_g(\boldsymbol{T}_{1j})\phi_l(\boldsymbol{T}_{1j}) \mid \{\zeta_g : g \geq 1\}\right\} \\ &\quad \times \mathbb{E}\left\{\sum_{g=1}^{\infty}\zeta_g\phi_g(\boldsymbol{T}_{1j'})\phi_l(\boldsymbol{T}_{1j'}) \mid \{\zeta_g : g \geq 1\}\right\} - \mathbb{E}\zeta_l\zeta_h\Bigg]^2 \\ &= \mathbb{E}(\zeta_l\zeta_h - \mathbb{E}\zeta_l\zeta_h)^2 \leq \mathbb{E}\left(\zeta_l^2\zeta_h^2\right). \end{aligned}$$

Putting together all these cases leads to

$$\begin{aligned} &\sum_{l,h=1}^{\infty}\frac{\mu_l\mu_h}{\nu_{lh}}\mathbb{E}\left\{\sum_{j\neq j'}^{m} e_{1jj'}\phi_l(\boldsymbol{T}_{1j})\phi_h(\boldsymbol{T}_{1j'})\right\}^2 \\ &\leq \sum_{l,h=1}^{\infty}\frac{\mu_l\mu_h}{\nu_{lh}}\left\{m(m-1)c\sigma_{\psi_1}^2 + 3m(m-1)(m-2)c\sigma_{\psi_1}^2 + m(m-1)(m-2)(m-3)\mathbb{E}\left(\zeta_l^2\zeta_h^2\right)\right\} \\ &\leq c\left\{m^3c\sigma_{\psi_1}^2\sum_{l,h=1}^{\infty}\frac{\mu_l\mu_h}{\nu_{lh}} + m^4\sum_{l,h=1}^{\infty}\frac{\mu_l\mu_h}{\nu_{lh}}\mathbb{E}\left(\zeta_l^2\zeta_h^2\right)\right\} \\ &\leq c\left\{m^3c\sigma_{\psi_1}^2\sum_{l,h=1}^{\infty}\min\{t^2, \mu_l\mu_h\} + m^4t^2\sum_{l,h=1}^{\infty}\mathbb{E}\left(\zeta_l^2\zeta_h^2\right)\right\}. \end{aligned}$$

Since $\sum_{l,h=1}^{\infty}\mathbb{E}\left(\zeta_l^2\zeta_h^2\right)=\mathbb{E}(X^4(\boldsymbol{T}))=L<\infty$,

$$\mathbb{E}\left\{\tilde{Z}_{n,m}(e,t;\mathcal{G}')\right\}^2\leq c_\eta\left(\frac{1}{nm}\sum_{l,h=1}^{\infty}\min\{t^2,\mu_l\mu_h\}+\frac{t^2}{n}\right).$$

□

Next we derive the following concentration inequality for $\tilde{Z}_{n,m}(e,t;\mathcal{G}')$.

Lemma S2. *There exists a universal constant $c>1$ and a constant $c_1>0$ depending on b and σ_{ψ_1}, such that with probability at least $1-\exp(-cnt^2/\log n)$, we have*

$$\tilde{Z}_{n,m}(e,t;\mathcal{G}')\leq c\left\{\mathbb{E}\tilde{Z}_{n,m}(e,t;\mathcal{G}')+c_1t^2\right\}.$$

Proof. Write $\boldsymbol{e}_i=\left\{e_{ijj'}:j=1,...,m\right\}$, $\boldsymbol{T}_i=\{\boldsymbol{T}_{ij}:j=1,...,m\}$ and

$$f(\boldsymbol{e}_i,\boldsymbol{T}_i)=\frac{1}{m(m-1)}\sum_{j\neq j'}^{m}e_{ijj'}g(\boldsymbol{T}_{ij},\boldsymbol{T}_{ij'}).$$

Note that $\mathbb{E}(f(\boldsymbol{e}_i,\boldsymbol{T}_i))=0$. We adopt the Adamczak bound (Theorem 4 in Adamczak et al., 2008; Koltchinskii, 2011) to establish a concentration inequality for the unbounded class $\mathcal{F}=\{f:g\in\mathcal{G}',\|g\|_2\leq t\}$. To this end, we need to bound a variance term $\sigma^2(\mathcal{F}):=\sup_{f\in\mathcal{F}}\mathbb{E}(f^2(\boldsymbol{e}_1,\boldsymbol{T}_1))$ and the sub-exponential norm of the envelope function F of the class $\mathcal{F}$. For the variance term,

$$\begin{aligned}
\sigma^2(\mathcal{F}):=&\sup_{\|g\|_{\mathcal{G}}\leq1,\|g\|_2\leq t}\mathbb{E}f^2(\boldsymbol{e}_1,\boldsymbol{T}_1))\\
=&\frac{1}{m^2(m-1)^2}\sup_{\|g\|_{\mathcal{G}}\leq1,\|g\|_2\leq t}\mathbb{E}\left\{\sum_{j\neq j'}^{m}e_{1jj'}g(\boldsymbol{T}_{1j},\boldsymbol{T}_{1j'})\right\}^2\\
=&\frac{1}{m^2(m-1)^2}\sum_{j\neq j'}^{m}\sum_{k\neq k'}^{m}\sup_{\|g\|_{\mathcal{G}'}\leq1,\|g\|_2\leq t}\mathbb{E}(\mathbb{E}e_{1jj'}e_{1kk'}\mid\boldsymbol{T}_1)g(\boldsymbol{T}_{1j},\boldsymbol{T}_{1j'})g(\boldsymbol{T}_{ik},\boldsymbol{T}_{ik'})\\
\leq&\frac{c\sigma_{\psi_1}^2}{m^2(m-1)^2}\sum_{j\neq j'}^{m}\sum_{k\neq k'}^{m}\sup_{\|g\|_{\mathcal{G}}\leq1,\|g\|_2\leq t}\mathbb{E}g(\boldsymbol{T}_{1j},\boldsymbol{T}_{1j'})g(\boldsymbol{T}_{ik},\boldsymbol{T}_{ik'})\\
\leq&\frac{c\sigma_{\psi_1}^2}{m^2(m-1)^2}\sum_{j\neq j'}^{m}\sum_{k\neq k'}^{m}\sup_{\|g\|_{\mathcal{G}}\leq1,\|g\|_2\leq t}\left\{\mathbb{E}g^2(\boldsymbol{T}_{1j},\boldsymbol{T}_{1j'})\mathbb{E}g^2(\boldsymbol{T}_{ik},\boldsymbol{T}_{ik'})\right\}^{1/2}\\
\leq&c\sigma_{\psi_1}^2t^2.
\end{aligned}$$

As for the envelope,

$$\begin{aligned}
\|\max_{i=1,...,n}F(\boldsymbol{e}_i,\boldsymbol{T}_i)\|_{\psi_1}\leq&\,c\max_{i=1,...,n}\|F(\boldsymbol{e}_i,\boldsymbol{T}_i)\|_{\psi_1}(\log n)\\
\leq&\frac{cb}{m(m-1)}\|\sum_{j\neq j'}^{m}e_{ijj'}\|_{\psi_1}(\log n)\leq cb\sigma_{\psi_1}^2(\log n),
\end{aligned}$$

where the first inequality comes from Theorm 4 of Pisier (1983) and the second inequality results from $g(\boldsymbol{T}_{ij},\boldsymbol{T}_{ij'})\leq b$. The desired result then follows from Adamzack bound. □

By Lemmas S1 and S2, we are able to bound $\tilde{Z}_{n,m}(e,t;\mathcal{G}')$. Then, we relate $\hat{Z}_{n,m}(e,t;\mathcal{G}')$ with $\tilde{Z}_{n,m}(e,t;\mathcal{G}')$ by Lemma S3 below. Recall that $\kappa_{n,m}$ is the smallest positive real number κ that fulfills the following inequalities

$$cb^3 Q(\kappa/b) \le \kappa^2, \tag{S3}$$
$$32cbQ(\kappa) \le \kappa^2, \tag{S4}$$

where c is an universal constant that we do not specify and

$$Q(\kappa) = \left[\frac{1}{n(m-1)} \sum_{l,h=1}^{\infty} \min\left\{\kappa^2, \mu_l \mu_h\right\}\right]^{1/2}.$$

Note that $Q(\kappa)/\kappa \to \infty$ as $\kappa \to 0$. Also, $Q(\kappa)/\kappa$ is non-increasing in κ. Dividing both sides in (S3) and (S4) by κ, the resulting right hand side is an identity function, which is continuous, strictly increasing and is zero when $\kappa = 0$. Therefore $\kappa_{n,m}$ exists.

Lemma S3. *We assume $t \ge \kappa_{n,m}$ for all the following cases. For any $g \in \mathcal{G}'$, there exist constants $M_1, M_2 > 2$, both depending on b, such that*

$$\left\{\|g\|_{n,m}^2 \le t^2\right\} \subseteq \left\{\|g\|_2^2 \le M_1 t^2\right\},$$

with probability at least $1 - \exp(-cnm\kappa_{n,m}^2 + \log m)$, and

$$\left\{\|g\|_2^2 \le t^2\right\} \subseteq \left\{\|g\|_{n,m}^2 \le M_2 t^2\right\},$$

with probability at least $1 - \exp(-cnm\kappa_{n,m}^2 + \log m)$. Additionally, we have

$$\|g\|_2^2 - \|g\|_{n,m}^2 \le \frac{1}{2}\|g\|_2^2,$$

holds for all $g \in \mathcal{G}'$ such that $\|g\|_2^2 > t^2$, with probability at least $1 - \exp(-c_p nmt^2 + \log m)$ where c_p is a constant depending on b.

Proof. For $1 \le j, j' \le m$, we call (j, j') a pair formed by individuals j and j'. When m is even, by Lemma S4, we are able to partition the collection $\mathcal{P} = \{(j,j') : 1 \le j < j' \le m\}$ into $(m-1)$ groups $G_1, ..., G_{m-1}$, such that $G_k \cap G_{k'} = \emptyset$ for $k \ne k'$, $\mathcal{P} = \bigcup_{k=1}^{m-1} G_k$, $\mathrm{card}(G_k) = m/2$ for all k, and $\mathrm{card}(\{(j,j') \in G_k : j = \tilde{j} \text{ or } j' = \tilde{j}\}) = 1$ for all $\tilde{j}$ and k (i.e., no individual occurs more than one time within a group), where $\mathrm{card}(A)$ denotes the cardinality of a set A. Therefore it is easy to see that the location pairs in $\{(T_{ij}, T_{ij'}) : (j,j') \in G_k\}$ are independent for any fixed k. As an illustration, suppose $m = 4$. Following the construction rule in Lemma S4, we obtain three groups $G_1 = \{(1,4),(2,3)\}$, $G_2 = \{(1,2),(3,4)\}$ and $G_3 = \{(1,3),(2,4)\}$.

Consider the case when m is even. Take $f_{G_k}(\boldsymbol{T}) = \frac{2}{nm}\sum_{i=1}^{n}\sum_{(j,j')\in G_k} g^2(\boldsymbol{T}_{ij}, \boldsymbol{T}_{ij'})$, $k = 1, ..., m-1$. Note that the $nm/2$ summands $g^2(\boldsymbol{T}_{ij}, \boldsymbol{T}_{ij'})$ in $f_{G_k}(\boldsymbol{T})$ all have expectation $\|g\|_2^2$, and are independent due to the above grouping property. To relate $\|g\|_2^2$ and $f_{G_k}(\boldsymbol{T})$, we can apply Theorem 3.3 in Bartlett et al. (2005).

Take $R_{n,m}(t; G_k, \mathcal{G}') = \frac{2}{nm} \sup_{\{g \in \mathcal{G}': \|g\|_2 \le t\}} |\sum_{i=1}^n \sum_{(j,j') \in G_k} \sigma_{ijj'} g^2(\boldsymbol{T}_{ij}, \boldsymbol{T}_{ij'})|$ to be the corresponding empirical local Rademacher complexity. By the well-known contraction inequality and Lemma 42 in Mendelson (2002), it is simple to show that with some universal constant c,

$$\begin{aligned}\mathbb{E} R_{n,m}(t; G_k, \mathcal{G}') &\le 2b \frac{2}{nm} \mathbb{E} \left\{ \sup_{g \in \mathcal{G}', \|g\|_2 \le t} \left| \sum_{(j,j') \in G_k} \sigma_{ijj'} g(\boldsymbol{T}_{ij}, \boldsymbol{T}_{ij'}) \right| \right\} \\ &\le cb \left(\frac{1}{nm} \sum_{l,h=1}^{\infty} \min\{t^2, \mu_l \mu_h\} \right)^{1/2} \le cbQ(t)\end{aligned}$$

Note that for $(j, j') \in \mathcal{G}_k$,

$$\mathrm{Var}\{g^2(\boldsymbol{T}_{ij}, \boldsymbol{T}_{ij'})\} \le \mathbb{E}\{g^4(\boldsymbol{T}_{ij}, \boldsymbol{T}_{ij'})\} \le b^2 \|g\|_2^2 \le b^2 t^2.$$

In Theorem 3.3 in Bartlett et al. (2005), we can take $T(g) = b^2 \|g\|_2^2$, $B = b^2$ and $\psi(r) = cb^3 Q(r^{1/2}/b)$. We then verify a condition in Theorem 3.3 in Bartlett et al. (2005). For any $t > 0$,

$$b^2 \mathbb{E} R_{n,m}(t; G_k, \mathcal{G}') = \frac{2b^2}{nm} \mathbb{E} \left\{ \sup_{g \in \mathcal{G}', T(g) \le b^2 t^2} \left| \sum_{(j,j') \in G_k} \sigma_{ijj'} g^2(\boldsymbol{T}_{ij}, \boldsymbol{T}_{ij'}) \right| \right\} \le cb^3 Q(t),$$

where the desired condition follows from taking $r = b^2 t^2$. From the definition (S3) of $\kappa_{n,m}$, we can see that $\kappa_{n,m}^2$ is larger than the fixed point of ψ (i.e., the solution of $\psi(r) = r$). Theorem 3.3 in Bartlett et al. (2005) shows that

$$\|g\|_2^2 \le 2 f_{G_k}(\boldsymbol{T}) + \frac{1408}{b^2} \kappa_{n,m}^2 + 2(11b^2 + 52b^2) \kappa_{n,m}^2 = 2 f_{G_k}(\boldsymbol{T}) + \left(\frac{1408}{b^2} + 126 b^2 \right) \kappa_{n,m}^2,$$

holds for all $g \in \mathcal{G}'$, with probability at least $1 - \exp(-nm\kappa_{n,m}^2)$. Also,

$$f_{G_k}(\boldsymbol{T}) \le 2\|g\|_2^2 + \frac{704}{b^2} \kappa_{n,m}^2 + 2(11b^2 + 26b^2) \kappa_{n,m}^2 = 2\|g\|_2^2 + \left(\frac{704}{b^2} + 74 b^2 \right) \kappa_{n,m}^2,$$

holds for all $g \in \mathcal{G}'$, with probability at least $1 - \exp(-nm\kappa_{n,m}^2)$.

Recall that $\|g\|_{n,m}^2 = \frac{1}{m-1} \sum_{i=1}^{m-1} f_{G_k}(\boldsymbol{T})$. We proceed by taking union bounds of the probability statements derived above, over $f_{G_1}, ..., f_{G_{m-1}}$. If $t \ge \kappa_{n,m}$,

$$\|g\|_2^2 \le 2\|g\|_{n,m}^2 + \left(\frac{1408}{b^2} + 126 b^2 \right) \kappa_{n,m}^2 \le M_1 t^2,$$

holds for all $g \in \mathcal{G}'$ such that $\|g\|_{n,m}^2 \le t^2$, with probability at least $1 - (m-1)\exp(-nm\kappa_{n,m}^2)$. Also, if $t \ge \kappa_{n,m}$,

$$\|g\|_{n,m}^2 \le 2\|g\|_2^2 + \left(\frac{704}{b^2} + 74 b^2 \right) \kappa_{n,m}^2 \le M_2 t^2,$$

holds for all $g \in \mathcal{G}'$ such that $\|g\|_2^2 \le t^2$, with probability at least $1-(m-1)\exp(-nm\kappa_{n,m}^2)$. Here $M_1, M_2 > 2$ are constants that depend on b.

Now, we focus on $\|g\|_2^2 > t^2$. By applying Theorem 2.1 in Bartlett et al. (2005), we obtain the following inequality

$$\|g\|_2^2 - f_{G_k}(\boldsymbol{T}) \le 0.5\|g\|_2^2,$$

holds for all $g \in \mathcal{G}'$ such that $\|g\|_2^2 > t^2$, with probability at least $1-\exp(-(mn/64b^2)t^2$. Take a union bound over $(m-1)$ groups, we will have

$$\|g\|_2^2 - \|g\|_{n,m}^2 \le 0.5\|g\|_2^2,$$

holds for all $g \in \mathcal{G}'$ such that $\|g\|_2^2 > t^2$, with probability at least $1-(m-1)\exp(-(mn/64b^2)t^2)$.

When m is odd, $\{(j,j') : 1 \le j < j' \le m-1\}$ can be decomposed into $(m-2)$ groups $(G_1, \ldots, G_{m-2})$ as described before, since $m-1$ is even. The remaining pairs are $\{(j,m) : j = 1, 2, ..., m-1\}$ which are not independent.

$$\|g\|_{n,m}^2 = \frac{m-2}{m}\frac{1}{(m-2)}\left\{\sum_{k=1}^{m-2}\sum_{i=1}^{n}\frac{2}{n(m-1)}\sum_{(j,j')\in G_k} g^2(\boldsymbol{T}_{ij}, \boldsymbol{T}_{ij'})\right\} + \frac{2}{m(m-1)}\sum_{j=1}^{m-1}\frac{1}{n}\sum_{i=1}^{n} g^2(\boldsymbol{T}_{ij}, \boldsymbol{T}_{im}) \tag{S5}$$

In the odd-m setting, we define $f_{G_k}(\boldsymbol{T}) = \frac{2}{n(m-1)}\sum_{i=1}^{n}\sum_{(j,j')\in G_k} g^2(\boldsymbol{T}_{ij}, \boldsymbol{T}_{ij'})$. We can apply the similar arguments derived for the even case (with m replaced by $m-1$). Therefore, we focus on the new term, which is the second term in (S5). First, we study $V_j(\boldsymbol{T}) = \frac{1}{n}\sum_{i=1}^{n} g^2(\boldsymbol{T}_{ij}, \boldsymbol{T}_{im})$ for a fixed $1 \le j \le m-1$. Note that $\mathbb{E}\{g^2(\boldsymbol{T}_{ij}, \boldsymbol{T}_{im})\} = \|g\|_2^2$ and the summands in $V_j(\boldsymbol{T})$ are independent.

We still apply Theorem 3.3 in Bartlett et al. (2005). The local Rademacher complexity becomes

$$R(t;\mathcal{G}') = \mathbb{E}\left\{\frac{1}{n}\sup_{g\in\mathcal{G}', \|g\|_2\le t}\sum_{i=1}^{n} g^2(\boldsymbol{T}_{ij}, \boldsymbol{T}_{im})\right\} \le cb\left(\frac{1}{n}\sum_{l,h=1}^{\infty}\min\{t^2, \mu_l\mu_h\}\right)^{1/2}.$$

Take κ_n' to be the smallest positive real number κ that satisfies

$$cb^3\left(\frac{1}{n}\sum_{l,h=1}^{\infty}\min\{(\kappa/b)^2, \mu_l\mu_h\}\right)^{1/2} \le \kappa^2.$$

By Theorem 3.3 in Bartlett et al. (2005), it can be shown that

$$\|g\|_2^2 \le 2V_j(\boldsymbol{T}) + \frac{1408}{b^2}\kappa_n'^2 + 2(11b^2 + 52b^2)m\kappa_{n,m}^2$$
$$\|g\|_2^2/m \le 2V_j(\boldsymbol{T})/m + \frac{1408}{b^2}\kappa_n'^2/m + 2(11b^2 + 52b^2)\kappa_{n,m}^2$$

holds for all $g \in \mathcal{G}'$, with probability at least $1-\exp(-nm\kappa_{n,m}^2)$. Also,

$$V_j(\boldsymbol{T})/m \le 2\|g\|_2^2/m + \frac{704}{b^2}\kappa_n'^2/m + 2(11(b^2) + 26b^2)\kappa_{n,m}^2,$$

holds for all $g \in \mathcal{G}'$, with probability at least $1 - \exp(-nm\kappa_{n,m}^2)$.

Now, we take a union bound, and then combine it with the result for the first term in (S5). Since $\kappa_n'^2/m \leq \kappa_{n,m}^2$, we derive the following with assumption $t \geq \kappa_{n,m}^2$:

$$\|g\|_2^2 \leq 2\|g\|_{n,m}^2 + \frac{1408}{b^2}\kappa_{n,m}^2\left(\frac{m-2}{m}+2\right) + 2(11b^2+52b^2)\kappa_{n,m}^2\left(\frac{m-2}{m}+2\right) \leq M_1 t^2$$

holds for all $g \in \mathcal{G}'$ such that $\|g\|_2^2 \leq t^2$, with probability at least $1 - (m-2+2(m-1))\exp(-nm\kappa_{n,m}^2)$.

$$\|g\|_{n,m}^2 \leq 2\|g\|_2^2 + \frac{704}{b^2}\kappa_{n,m}^2\left(\frac{m-2}{m}+2\right) + 2(11(b^2)+26b^2)\left(\frac{m-2}{m}+2\right)\kappa_{n,m}^2 \leq M_2 t^2$$

holds for all $g \in \mathcal{G}'$ such that $\|g\|_2^2 < t^2$, with probability at least $1 - (m-2+2(m-1))\exp(-nm\kappa_{n,m}^2)$. Here M_1 and M_2 are some constants that depend on b.

With similar argument, we will be able to derive for the odd case,

$$\|g\|_2^2 - \|g\|_{n,m}^2 \leq 0.5\|g\|_2^2,$$

holds for all $g \in \mathcal{G}'$ such that $\|g\|_2^2 > t^2$, with probability at least $1 - (m-2+2(m-1))\exp(-c_p nmt^2)$ for some constant $c_p = c_p(1/b)$.

□

With Lemmas S1, S2 and S3, we are now ready to prove Theorem 2. Recall the definition of $\eta_{n,m}$ and $\xi_{n,m}$. The term $\eta_{n,m}$ is defined as the smallest positive value η such that

$$\left(\frac{c_\eta}{nm}\sum_{l,h=1}^{\infty}\min\{\eta^2, \mu_l\mu_h\} + \frac{\eta^2}{n}\right)^{1/2} \leq \eta^2,$$

where $c_\eta > 0$ is a constant defined in Lemma S1. By similar arguments for the existence of $\kappa_{n,m}$, we can show that $\eta_{n,m}$ exists. By Lemma S1, we can show that $\mathbb{E}\tilde{Z}_{n,m}(e,t;\mathcal{G}') \leq \sqrt{\mathbb{E}[\{\tilde{Z}_{n,m}(e,t;\mathcal{G}')\}^2]} \leq t^2$ for $t \geq \eta_{n,m}$.

Take $\xi_{n,m} = \min\left\{\max\{\eta_{n,m}, \kappa_{n,m}\}, \left(\frac{\log n}{n}\right)^{1/2}\right\}$. We include the term $(\log n/n)^{1/2}$ mainly due to the unboundedness of $\{e_{ijj'}\}$, which leads to the application of Adamzack bound (Lemma S2) instead of simpler forms of Talagrand's concentration inequality.

Proof for Theorem 2. First, we study the crucial term

$$\hat{Z}_{n,m}(e,b;\mathcal{G}') = \sup_{\{g\in\mathcal{G}':\|g\|_{n,m}\leq b\}}\left|\frac{1}{nm(m-1)}\sum_{i=1}^{n}\sum_{j\neq j'}e_{ijj'}g(\boldsymbol{T}_{ij},\boldsymbol{T}_{ij'})\right|,$$

which is bounded by the maximum of $\hat{Z}_{n,m}(e,\xi_{n,m};\mathcal{G}')$ and

$$\sup_{\{g\in\mathcal{G}':\|g\|_{n,m}>\xi_{n,m}\}}\left|\frac{1}{nm(m-1)}\sum_{i=1}^{n}\sum_{j\neq j'}e_{ijj'}g(\boldsymbol{T}_{ij},\boldsymbol{T}_{ij'})\right|, \tag{S6}$$

so it suffices to study the rates of convergence of these two terms.

For the rate of the maximum of $\hat{Z}_{n,m}(e, \xi_{n,m}; \mathcal{G}')$, by Lemmas S1, S2 and S3, we can show that with probability at least $1 - \exp(-cn\xi_{n,m}^2/\log n)$ for some universal constant c:

$$\begin{aligned}\hat{Z}_{n,m}(e, \xi_{n,m}; \mathcal{G}') &\le \tilde{Z}_{n,m}(e, \sqrt{M_1}\xi_{n,m}; \mathcal{G}') \\ &\le c\left\{\mathbb{E}\tilde{Z}_{n,m}(e, \sqrt{M_1}\xi_{n,m}; \mathcal{G}') + c_1 M_1 \xi_{n,m}^2\right\} \\ &\le c\left\{M_1\xi_{n,m}^2 + c_1 M_1 \xi_{n,m}^2\right\} \le R\xi_{n,m}^2,\end{aligned}$$

where $R = cM_1(1 + c_1)$ and, the first, second and last inequalities are due to Lemmas S3, S2 and S1 respectively.

For the rate of the second term in (S6), we first prove the following result. For any $r > \xi_{n,m}$, with probability at least $1 - \exp(-cn\xi_{n,m}^2/\log n)$, we have

$$\begin{aligned}\hat{Z}_{n,m}(e, r; \mathcal{G}') &= \frac{r}{\xi_{n,m}} \sup_{\{g\in\mathcal{G}': \|g\|_{\mathcal{G}} \le \frac{\xi_{n,m}}{r}, \|g\|_{n,m} \le \xi_{n,m}\}} \left| \frac{1}{nm(m-1)} \sum_{i=1}^{n} \sum_{j\neq j'} e_{ijj'} g(\boldsymbol{T}_{ij}, \boldsymbol{T}_{ij'}) \right| \\ &\le \frac{r}{\xi_{n,m}} \hat{Z}_{n,m}(e, \xi_{n,m}; \mathcal{G}') \le \frac{r}{\xi_{n,m}} R\xi_{n,m}^2 = Rr\xi_{n,m}. \end{aligned} \tag{S7}$$

For $b > \xi_{n,m}$, a direct application of the above result with $r = b$ does not provide the increment with respect to the empirical norm, and so we apply a peeling argument.

Set $S_l := \{g \in \mathcal{G}' : 2^{l-1}\xi_{n,m} \le \|g\|_{n,m} \le 2^l \xi_{n,m}\}$, $l = 1, \ldots, L$, where $L = \log_2(b/\xi_{n,m})$.

$$\begin{aligned}&\mathbb{P}\left(\sup_{\{g\in\mathcal{G}': \|g\|_{n,m} > \xi_{n,m}\}} \frac{\left|\frac{1}{nm(m-1)} \sum_{i=1}^{n} \sum_{j\neq j'} e_{ijj'} g(\boldsymbol{T}_{ij}, \boldsymbol{T}_{ij'})\right|}{\|g\|_{n,m}} > 2R\xi_{n,m} \right) \\ &\le \sum_{l=1}^{L} \mathbb{P}\left(\sup_{g\in S_l} \frac{\left|\frac{1}{nm(m-1)} \sum_{i=1}^{n} \sum_{j\neq j'} e_{ijj'} g(\boldsymbol{T}_{ij}, \boldsymbol{T}_{ij'})\right|}{\|g\|_{n,m}} > 2R\xi_{n,m} \right) \\ &\le \sum_{l=1}^{L} \mathbb{P}\left(\hat{Z}_{n,m}(e, 2^l\xi_{n,m}; \mathcal{G}') > 2R2^{l-1}\xi_{n,m}^2 \right) \\ &\le L \exp\left(-c\frac{n}{\log n}\xi_{n,m}^2 \right) \\ &\le \exp\left(-c\frac{n}{\log n}\xi_{n,m}^2 \right),\end{aligned}$$

where the second last inequality results from (S7) by taking $r = 2^l\xi_{n,m}$, and the universal constant c in the last two lines could be different. For the last inequality, as long as $0 \le (\log(\log(1/\xi_{n,m})))/\{\frac{n}{\log n}\xi_{n,m}^2\} \le 1$ such a universal constant c exists.

Therefore, we have

$$\langle e, g\rangle_{n,m} \le R(\xi_{n,m}^2 + 2\|g\|_{n,m}\xi_{n,m}),$$

for every $g \in \mathcal{G}_s \subset \mathcal{G}'$, with probability at least $1 - \exp(-cn\xi_{n,m}^2/\log n)$. With the same probability, we have

$$\langle e, \Delta\rangle_{n,m} \le R\xi_{n,m}^2 \left\{ I(\hat{\Gamma}) + I(\Gamma_0) \right\} + 2R\xi_{n,m}\|\Delta\|_{n,m}. \tag{S8}$$

In below, we condition on the event (S8). From the basic inequality (S1), with $\lambda = c_\lambda \xi_{n,m}^2$ such that $c_\lambda > 2R$,

$$\|\Delta\|_{n,m}^2 \leq 2\langle e, \Delta\rangle_{n,m} + \lambda(I(\Gamma_0) - I(\hat{\Gamma})),$$
$$\|\Delta\|_{n,m}^2 \leq 2\lambda I(\Gamma_0) + 4R\xi_{n,m}\|\Delta\|_{n,m}.$$

Then we have

$$\|\Delta\|_{n,m} \leq \{2c_\lambda I(\Gamma_0)\}^{\frac{1}{2}}\, \xi_{n,m} + 4R\xi_{n,m}$$

and the proof is complete by taking $L_1 = 2R$. □

Next, we are ready to bound the L^2 norm $\|\Delta\|_2$ for $\Delta = \hat{\Gamma} - \Gamma_0$ obtained by (7).

Proof of Theorem 3. From Lemma S3, we can see that $\|g\|_2^2 \leq 2\|g\|_{n,m}^2 + \xi_{n,m}^2$, for all $g \in \mathcal{G}'$ with probability at least $1 - \exp(-c_p\xi_{n,m}^2)$ for some universal constant $c_p = c_p(1/b)$. So with the same probability we have $\|\Delta\|_2 \leq 2^{1/2}\|\Delta\|_{n,m} + \xi_{n,m}\left\{I(\hat{\Gamma}) + I(\Gamma_0)\right\}$. In terms of Lemma S5, we are able to bound the regularization term $I(\hat{\Gamma})$ by a constant L_2, so finally we get

$$\begin{aligned}\|\Delta\|_2 &\leq 2^{\frac{1}{2}}\left[\{2c_\lambda I(\Gamma_0)\}^{\frac{1}{2}} + 4R\right]\xi_{n,m} + \{R_2 + I(\Gamma_0)\}\,\xi_{n,m} \\ &\leq \left[2\{c_\lambda I(\Gamma_0)\}^{\frac{1}{2}} + 4(2)^{\frac{1}{2}}R + R_2 + I(\Gamma_0)\right]\xi_{n,m}.\end{aligned}$$

By taking $L_2 = 4(2)^{1/2}R + R_2 + I(\Gamma_0)$, the proof is complete. □

Proof of Corollary 1. By Lemma S6, the tensor product eigenvalue sequence has decay $\mu_l \asymp (l^{-2\alpha}(\log l)^{2\alpha(2p-1)})$ as $l \to \infty$.

By the definitions of $\kappa_{n,m}$ and $\eta_{n,m}$, when $m = \mathcal{O}\left(n^{1/(2\alpha)}(\log n)^{2p-2-1/(2\alpha)}\right)$, they are all of the same order, and so is $\xi_{n,m}$. By Lemma S7, we can see that $\xi_{n,m} \asymp (nm)^{2\alpha/(1+2\alpha)}(\log nm)^{2\alpha(2p-1)/(2\alpha+1)}$. When $n^{1/(2\alpha)}(\log n)^{2p-2-\frac{1}{2\alpha}} = \mathcal{O}(m)$, $\log n/n$ will be the dominant term. From Theorems 2 and 3, we can see that $\|\hat{\Gamma} - \Gamma_0\|_{n,m}^2$ and $\|\hat{\Gamma} - \Gamma_0\|_2^2$ are both of the same order. Overall, we have

$$\|\hat{\Gamma} - \Gamma_0\|_{n,m}^2, \|\hat{\Gamma} - \Gamma_0\|_2^2 = \mathcal{O}_{\mathrm{p}}\left((nm)^{-\frac{2\alpha}{1+2\alpha}}(\log nm)^{\frac{2\alpha(2p-1)}{2\alpha+1}} + \frac{\log n}{n}\right).$$

□

S1.3 Auxiliary Lemmas

Lemma S4. *When m is even, we can decompose any collection of individual index pairs $\{(j, j') : 1 \leq j < j' \leq m\}$ into $(m-1)$ groups such that within each group, there are $m/2$ pairs and no repeated individuals.*

Proof. First, we consider to construct a matrix $\boldsymbol{G} \in \mathbb{R}^{m\times m}$ that satisfies following conditions:
1. All the diagonal entries are zero;
2. Every row and every column is a permutation of sequence $\{0, 1, 2, ..., (m-1)\}$;
3. It is symmetric.

To begin with, we consider the cycle $cyc = \{1, 2, ..., (m-1)\}$ and construct a sub-matrix $\boldsymbol{G}_{sub} \in \mathbb{R}^{(m-1)\times(m-1)}$ from it. For $i-$th row of $\boldsymbol{G}_{sub}$, we set it to be a sequence that starts with i in cyc and ends until it reaches $(m-1)$ elements. For example, the first row will be $[1, 2, ..., (m-1)]$, the second row will be $[2, 3, ..., (m-1), 1]$, and so on. Take the first $(m-1)$ rows and first $(m-1)$ columns of $\boldsymbol{G}$ to be $\boldsymbol{G}_{sub}$ and fill last row and last column of $\boldsymbol{G}$ with zeros. Then obviously $\boldsymbol{G}$ fulfills Conditions 2 and 3.

To fulfill Condition 1, set $\boldsymbol{G}_{i,m}$ and $\boldsymbol{G}_{m,i}$ to be $\boldsymbol{G}_{ii}$ and then set $\boldsymbol{G}_{ii} = 0$ for $i = 1, ..., (m-1)$. By this operation, it's easy to see that for first $(m-1)$ rows and first $(m-1)$ columns, they are still permutations of sequence $\{0, 1, 2, ..., (m-1)\}$ and symmetrization of $\boldsymbol{G}$ is not violated. It remains to prove that last row and last column are also a permutation of the sequence, which is equivalent to proving the diagonal part of $\boldsymbol{G}_{sub}$ is a permutation. In fact $\boldsymbol{G}_{sub(i,i)}$ is $(2i-1)$-th element of cycle cyc, $i = 1, 2, ..., (m-1)$. Since m is even, diagonal parts of $\boldsymbol{G}_{sub}$ will cover the whole sequence $\{1, 2, ..., (m-1)\}$.

So for every pair (j, j'), $1 \le j < j' \le m$, we can assign it to Group G_k where $k = \boldsymbol{G}_{j,j'}$. In this way, we decompose the collection $\{(j, j') : 1 \le j < j' \le m\}$ into $(m-1)$ groups where each group contains $m/2$ elements and within one group, there is no repeated individual. □

Lemma S5. *Under the same assumptions as Theorem 2, if $\lambda = c_\lambda \xi_n^2$ with some constant $c_\lambda > 2R$, then there exists a constant R_2 depending on $I(\Gamma_0)$, R and c_λ, such that with probability at least $1 - \exp(c\frac{n}{\log n}\xi_{n,m}^2)$, we have*

$$I(\hat{\Gamma}) \le R_2.$$

Proof. From the basic inequality (S1), we have

$$\|\Delta\|_{n,m}^2 + \lambda I(\hat{\Gamma}) \le 2\langle e, \Delta\rangle + \lambda I(\Gamma_0), \tag{S9}$$

$$\lambda I(\hat{\Gamma}) \le 2\langle e, \Delta\rangle + \lambda I(\Gamma_0). \tag{S10}$$

From Theorem 2, we know that

$$\langle e, \Delta\rangle \le R\xi_{n,m}^2 \left\{ I(\hat{\Gamma}) + I(\Gamma_0) \right\} + 2R\xi_{n,m}\|\Delta\|_{n,m}, \tag{S11}$$

and

$$\|\Delta\|_{n,m} \le \{2c_\lambda I(\Gamma_0)\}^{\frac{1}{2}} \xi_{n,m} + 4R\xi_{n,m}. \tag{S12}$$

Therefore, plug (S12) into (S12),

$$\langle e, \Delta\rangle \le \left[R \left\{ I(\hat{\Gamma}) + I(\Gamma_0) \right\} + 2R \{2c_\lambda I(\Gamma_0)\}^{\frac{1}{2}} + 8R^2 \right] \xi_{n,m}^2.$$

By plugging in(S10), we have

$$(c_\lambda - 2R) I(\hat{\Gamma}) \le 2RI(\Gamma_0) + 4R \{2c_\lambda I(\Gamma_0)\}^{\frac{1}{2}} + 16R^2 + c_\lambda I(\Gamma_0).$$

Therefore, there exists a constant L_2, such that

$$I(\hat{\Gamma}) \le \frac{2RI(\Gamma_0) + 4R \{2c_\lambda I(\Gamma_0)\}^{\frac{1}{2}} + 16R^2 + c_\lambda I(\Gamma_0)}{c_\lambda - 2R} \le R_2.$$

□

Lemma S6. *Suppose $K_1(\cdot,\cdot) = K_2(\cdot,\cdot) = \dots K_p(\cdot,\cdot)$, then $\mathcal{H}_1 = \mathcal{H}_2 = \dots = \mathcal{H}_p$. If eigenvalues of K_k has decay $\mu_n^{(k)} \asymp (n^{-s})$ for some constant s. Then eigenvalues of the reproducing kernel for tensor product $\bigotimes_{k=1}^{p} \mathcal{H}_k \otimes \bigotimes_{k=1}^{p} \mathcal{H}_k$ will have decay $\mu_n \asymp (n^{-s}(\log n)^{s(2p-1)})$*

Proof. A direct application of Theorem 1 (Krieg, 2018) completes the proof. □

Lemma S7. *Take t to be the solution of the equality*

$$\frac{1}{\sqrt{nm}}\left(\sum_{h=1}^{\infty}\min\left\{t^2,\mu_h\right\}\right)^{1/2} = t^2,$$

where $\mu_h \asymp (h^{-2\alpha}(\log l)^{2\alpha(2p-1)})$. Then as $n \to \infty$ and $m \to \infty$, the solution

$$t \asymp (nm)^{-\frac{\alpha}{1+2\alpha}}(\log nm)^{\frac{\alpha(2p-1)}{2\alpha+1}}.$$

Proof. Take $N = nm$, To find the order of t. We need to find l' such that $t^2 \asymp l'^{-2\alpha}(\log l')^{2\alpha(2p-1)}$. From some simple analysis, we could see that when $N \to \infty$, $t \to 0$ and $l' \to \infty$. Therefore, when $N \to \infty$ we have

$$\begin{aligned}
t &\asymp l'^{-\alpha}(\log l')^{\alpha(2p-1)},\\
\frac{1}{t} &\asymp l'^{\alpha}(\log l')^{-\alpha(2p-1)},\\
log(1/t) &\asymp \alpha\log l' - \alpha(2p-1)\log(\log(l')) \asymp \log l',\\
l' &\asymp t^{-1/\alpha}(\log(1/t))^{2p-1}.
\end{aligned}$$

It's easy to see that $t^2 l' \asymp (t)^{2-1/\alpha}(\log(1/t))^{2p-1}$, $\sum_{l\geq l'}\mu_l \asymp \mathcal{O}(l'^{-2\alpha+1}(\log l')^{2\alpha(2p-1)}) \asymp \mathcal{O}(t^{2-1/\alpha}(\log(1/t))^{2p-1})$.

So $\sum_{l\geq l'}\mu_l = \mathcal{O}(\xi_n^2 l')$, therefore

$$\begin{aligned}
\frac{\sqrt{t^2 l'}}{\sqrt{N}} &\asymp t^2,\\
N &\asymp (1/t)^{2+1/\alpha}(\log(1/t))^{2p-1},\\
\log N &\asymp (2+1/\alpha)log(1/\xi_n) + (2p-1)\log\log(1/t) \asymp \log(1/t),\\
1/t &\asymp N^{\frac{\alpha}{1+2\alpha}}(\log N)^{-\frac{\alpha(2p-1)}{2\alpha+1}},\\
t &\asymp N^{-\frac{\alpha}{1+2\alpha}}(\log N)^{\frac{\alpha(2p-1)}{2\alpha+1}}.
\end{aligned}$$

□

S2 Simulation

S2.1 Eigenfunctions in different simulation settings

We present three simulation settings in a table form (Tables S1, S2 and S3). In each table, rows correspond to basis functions for dimension 1 and columns correspond to basis functions

for dimension 2. Recall that for each dimension, we use $e_k(t) = \sqrt{2}\cos(k\pi t)$, $k = 1, 2, \ldots$ as basis. Then, for the cell with position row i and column j, it represents the two dimensional function $f_{ij}(s_1, s_2) = e_i(s_1)e_j(s_2)$. We use a positive integer k to indicate that this two dimensional function is the k-th eigenfunction. The details of the three settings are given as follows.

Table S1: Eigenfunctions for Setting 1

	e_1	e_2
e_1	1	2
e_2	3	5
e_3	4	6

Table S2: Eigenfunctions for Setting 2

	e_1	e_2	e_3	e_4
e_1	1	2	-	-
e_2	3	4	-	-
e_3	-	-	5	-
e_4	-	-	-	- 6

Table S3: Eigenfunctions for Setting 3

	e_1	e_2	e_3	e_4
e_1	-	1	-	-
e_2	2	-	-	-
e_3	-	-	3	-
e_4	-	-	-	- 4

Setting 1: $R = 6$, $r_1 = 3$, $r_2 = 2$. For dimension 1, we use e_1, e_2 and e_3 as our basis functions. For dimension 2, we use e_1 and e_2 as our basis functions. Let 6 eigenfunctions ψ_k be the tensor product of these one dimensional basis with eigenvalue decay $\lambda_k = 1/(k^2)$, $k = 1, 2, ..., 6$. Eigenfunctions can be expressed as $\psi_k(t_1, t_2) = e_i(t_1)e_j(t_2)$, where $k = 2(i-1)+j$ for $k = 1, 2, 3, 6$ and $\psi_4(s, t) = e_3(s)e_1(t)$, $\psi_5(t_1, t_2) = e_2(t_1)e_2(t_2)$. In this setting, $R = r_1 * r_2$, one-way basis are mostly shared among different eigenfunctions.

Setting 2: $R = 6$, $r_1 = r_2 = 4$. For both dimension 1 and dimension 2, we use e_i, $i = 1, .., 4$ as our basis functions. Let 6 eigenfunctions ψ_k with eigenvalue decay $\lambda_k = 1/(k^2)$, $k = 1, 2, ..., 6$. $\psi_k(t_1, t_2) = e_i(t_1)e_j(t_2)$, where $k = 2(i-1)+j$ for $k = 1, 2, 3$. $\psi_k(t_1, t_2) = e_{k-2}(t_1)e_{k-2}(t_2)$ for $k = 4, 5, 6$. In this setting, one-way basis are partly shared by different eigenfunctions.

Setting 3: $R = r_1 = r_2 = 4$. For both dimension 1 and dimension 2, we use e_i, $i = 1, .., 4$ as our basis functions. Let 4 eigenfunctions ψ_k with eigenvalue decay $\lambda_k = 1/(k^2)$, $k = 1, ..., 4$. $\psi_1(t_1, t_2) = e_1(t_1)e_2(t_2)$, $\psi_2(t_1, t_2) = e_2(t_1)e_1(t_2)$ and $\psi_k(t_1, t_2) = e_k(t_1)e_k(t_2)$ for $k = 3, 4$. In this case, one-way basis are not shared among different eigenfunctions.

S2.2 Additional simulation results for sparse design

The simulation results for the sparse design with sample size $n = 100$ are shown in Table S4.

Table S4: Simulation results for three Settings with the sparse design when sample size is 100 ($n = 100$): see description in Table 1.

Setting	m	σ		mOpCov	OpCov	ll-smooth	ll-smooth+
1	10	0.1	AISE	0.101 (5.47e-03)	0.122 (1.20e-02)	4.36 (2.28e+00)	1.702 (8.43e-01)
			$\bar{R}$	7.66	2.45	-	142.61
			$\bar{r}_1$, $\bar{r}_2$	5.07 , 5.04	_	_	_
		0.4	AISE	0.104 (5.62e-03)	0.12 (1.19e-02)	3.89 (1.78e+00)	0.989 (1.96e-01)
			$\bar{R}$	7.34	2.2	-	146.33
			$\bar{r}_1$, $\bar{r}_2$	4.84 , 4.82	_	_	_
	20	0.1	AISE	0.0661 (2.99e-03)	0.075 (3.49e-03)	3.93 (3.17e+00)	1.40 (9.80e-01)
			$\bar{R}$	7.84	3.02	-	249.95
			$\bar{r}_1$, $\bar{r}_2$	5.48 , 5.48	_	_	_
		0.4	AISE	0.0679 (3.06e-03)	0.0761 (3.34e-03)	0.468 (6.90e-02)	0.310 (2.32e-02)
			$\bar{R}$	7.51	2.83	-	205.675
			$\bar{r}_1$, $\bar{r}_2$	5.38 , 5.38	_	_	_
2	10	0.1	AISE	0.1 (5.38e-03)	0.113 (6.12e-03)	2.12 (6.23e-01)	0.826 (1.76e-01)
			$\bar{R}$	7.68	2.38	-	144.645
			$\bar{r}_1$, $\bar{r}_2$	5.53, 5.56	_	_	_
		0.4	AISE	0.102 (5.44e-03)	0.112 (5.64e-03)	4.18 (2.21e+00)	0.931 (1.76e-01)
			$\bar{R}$	7.34	2.22	-	146.855
			$\bar{r}_1$, $\bar{r}_2$	5.49, 5.49	_	_	_
	20	0.1	AISE	0.0637 (2.95e-03)	0.0706 (3.21e-03)	0.472 (8.01e-02)	0.304 (2.80e-02)
			$\bar{R}$	8.37	2.76	-	200.69
			$\bar{r}_1$, $\bar{r}_2$	5.81, 5.8	_	_	_
		0.4	AISE	0.0649 (3.06e-03)	0.0733 (3.30e-03)	0.484 (7.27e-02)	0.317 (2.53e-02)
			$\bar{R}$	8.24	2.78	-	206.16
			$\bar{r}_1$, $\bar{r}_2$	5.78, 5.78	_	_	_
3	10	0.1	AISE	0.105 (4.75e-03)	0.115 (7.58e-03)	24.1 (2.28e+01)	1.87 (1.19)
			$\bar{R}$	8.75	2.82	-	150.8
			$\bar{r}_1$, $\bar{r}_2$	5.26, 5.32	_	_	_
		0.4	AISE	0.11 (4.96e-03)	0.115 (8.33e-03)	26.2 (2.40e+01)	2.05
			$\bar{R}$	9.44	2.74	-	152.575
			$\bar{r}_1$, $\bar{r}_2$	5.37, 5.4	_	_	_
	20	0.1	AISE	0.0698 (2.74e-03)	0.0813 (4.63e-03)	0.614 (2.28e-01)	0.350 (8.35e-02)
			$\bar{R}$	6.63	3.24	-	210.515
			$\bar{r}_1$, $\bar{r}_2$	5.08, 5.14	_	_	_
		0.4	AISE	0.0721 (2.89e-03)	0.0859 (5.03e-03)	0.573 (1.74e-01)	0.344 (6.37e-02)
			$\bar{R}$	6.74	3.38	-	214.455
			$\bar{r}_1$, $\bar{r}_2$	5.11, 5.21	_	_	_

S2.3 Simulation results for regular design

For regular design, we selected 10 equally spaced points for each dimension and constructed a regular 10×10 grid ($m = 100$). We set sample size to be 50 ($n = 50$). Two different noise levels are considered, since regular design has dense observations, we pick $\sigma = 0.4$ to represent the low noise level and $\sigma = 0.8$ to represent the high noise level. Beside methods we

mentioned in sparse design, we also include an additional estimator from Wang and Huang (2017) (spatpca), which allows to perform multi-dimensional covariance function estimation with location-fixed observations into our comparisons. Results are showed in Table S5, Table S6 and Table S7.

Table S5: Results for Setting 1 on regular design: see description in Table 1.

σ		mOpCov	OpCov	ll-smooth	spatpca	ll-smooth+
0.40	AISE	0.0611 (4.37e-03)	0.0626 (4.10e-03)	0.0571 (3.71e-03)	0.0625 (4.37e-03)	0.057 (3.71e-03)
	$\hat{R}$	8.27	7.25	-	5.95	15.125 (0.10)
	$\hat{r}_1, \hat{r}_2$	6, 6	-	-	-	-
0.80	AISE	0.0629 (4.45e-03)	0.0676 (4.49e-03)	0.0643 (3.79e-03)	0.0738 (4.52e-03)	0.0639 (3.79e-03)
	$\hat{R}$	10.9	3.98	-	5.84	26.065 (0.087)
	r_1, r_2	6, 6	-	-	-	-

Table S6: Results for Setting 2 on regular design: see description in Table 1.

σ		mOpCov	OpCov	ll-smooth	spatpca	ll-smooth+
0.40	AISE	0.0602 (4.38e-03)	0.0641 (4.69e-03)	0.056 (3.71e-03)	0.0624 (4.37e-03)	0.0559 (3.71e-03)
	$\hat{R}$	8.09	7.22	-	4.21	14.135 (0.095)
	$\hat{r}_1, \hat{r}_2$	6, 6	-	-	-	-
0.80	AISE	0.062 (4.48e-03)	0.0659 (4.54e-03)	0.0631 (3.79e-03)	0.0724 (4.47e-03)	0.0627 (3.79e-03)
	$\hat{R}$	10.7	4.04	-	4.28	25.84 (0.0898)
	$\hat{r}_1, \hat{r}_2$	6, 6	-	-	-	-

Table S7: Results for Setting 3 on regular design: see description in Table 1.

σ		mOpCov	OpCov	ll-smooth	spatpca	ll-smooth+
0.40	AISE	0.0628 (4.22e-03)	0.0589 (4.34e-03)	0.0677 (3.92e-03)	0.0598 (4.17e-03)	0.0675 (3.92e-03)
	$\hat{R}$	5.66	14	-	3.52	18.74 (0.104)
	$\hat{r}_1, \hat{r}_2$	6, 6	-	-	-	-
0.80	AISE	0.0645 (4.05e-03)	0.0677 (4.48e-03)	0.0745 (3.94e-03)	0.0715 (4.20e-03)	0.07389 (3.94e-03)
	$\hat{R}$	7.7	13.1	-	2.93	29.485 (0.143)
	$\hat{r}_1, \hat{r}_2$	6, 6	-	-	-	-

References

Adamczak, R. et al. (2008). A tail inequality for suprema of unbounded empirical processes with applications to markov chains. *Electronic Journal of Probability 13*, 1000–1034.

Bartlett, P. L., O. Bousquet, and S. Mendelson (2005). Local rademacher complexities. *The Annals of Statistics 33*(4), 1497–1537.

Koltchinskii, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*, Volume 2033. Springer Science & Business Media.

Krieg, D. (2018). Tensor power sequences and the approximation of tensor product operators. *Journal of Complexity 44*, 30–51.

Mendelson, S. (2002). Geometric parameters of kernel machines. In *International Conference on Computational Learning Theory*, pp. 29–43. Springer.

Pisier, G. (1983). Some applications of the metric entropy condition to harmonic analysis. In *Banach Spaces, Harmonic Analysis, and Probability Theory*, pp. 123–154. Springer.

Wang, W.-T. and H.-C. Huang (2017). Regularized principal component analysis for spatial data. *Journal of Computational and Graphical Statistics 26*(1), 14–25.

Wong, R. K. W. and X. Zhang (2019). Nonparametric operator-regularized covariance function estimation for functional data. *Computational Statistics & Data Analysis 131*, 131–144.

Table S8: Simulation results for Setting 1 ($R = 6$, $r_1 = 3$, and $r_2 = 2$) with the sparse design. The AISE values with standard errors (SE) in parentheses are provided for the four covariance estimators in comparison, together with average two-way ranks ($\bar{R}$) for those estimators which can lead to rank reduction (i.e., mOpCov, OpCov, and ll-smooth+) and average one-way ranks ($\bar{r}_1$, $\bar{r}_2$) for mOpCov.

n	m	σ		mOpCov	OpCov	ll-smooth	ll-smooth+
100	10	0.1	AISE	0.101 (5.47e-03)	0.122 (1.20e-02)	4.36 (2.28e+00)	1.702 (8.43e-01)
			$\bar{R}$	7.66	2.45	-	142.61
			$\bar{r}_1$, $\bar{r}_2$	5.07 , 5.04	_	_	_
		0.4	AISE	0.104 (5.62e-03)	0.12 (1.19e-02)	3.89 (1.78e+00)	0.989 (1.96e-01)
			$\bar{R}$	7.34	2.2	-	146.33
			$\bar{r}_1$, $\bar{r}_2$	4.84 , 4.82	_	_	_
	20	0.1	AISE	0.0661 (2.99e-03)	0.075 (3.49e-03)	3.93 (3.17e+00)	1.40 (9.80e-01)
			$\bar{R}$	7.84	3.02	-	249.95
			$\bar{r}_1$, $\bar{r}_2$	5.48 , 5.48	_	_	_
		0.4	AISE	0.0679 (3.06e-03)	0.0761 (3.34e-03)	0.468 (6.90e-02)	0.310 (2.32e-02)
			$\bar{R}$	7.51	2.83	-	205.675
			$\bar{r}_1$, $\bar{r}_2$	5.38 , 5.38	_	_	_
200	10	0.1	AISE	0.053 (1.97e-03)	0.0632 (3.22e-03)	0.652 (1.92e-01)	0.337 (5.35e-02)
			$\bar{R}$	8.38	2.94	-	172.70
			$\bar{r}_1$, $\bar{r}_2$	5.4, 5.4	_	_	_
		0.4	AISE	0.0547 (2.01e-03)	0.0656 (2.72e-03)	0.714 (2.11e-01)	0.366 (5.96e-02)
			$\bar{R}$	9.16	2.84	-	177.3
			$\bar{r}_1$, $\bar{r}_2$	5.34, 5.32	_	_	_
	20	0.1	AISE	0.0343 (1.46e-03)	0.0421 (1.97e-03)	0.297 (1.39e-02)	0.206 (4.62e-03)
			$\bar{R}$	8.38	3.78	-	317.44
			$\bar{r}_1$, $\bar{r}_2$	5.84, 5.82	_	_	_
		0.4	AISE	0.0354 (1.52e-03)	0.044 (2.21e-03)	0.325 (1.58e-02)	0.223 (4.94e-03)
			$\bar{R}$	8.86	3.76	-	326.31
			$\bar{r}_1$, $\bar{r}_2$	5.83, 5.84	_	_	_

Table S9: Simulation results for Setting 2 ($R = 6$ and $r_1 = r_2 = 4$) with the sparse design: see description in Table S8.

n	m	σ		mOpCov	OpCov	ll-smooth	ll-smooth+
100	10	0.1	AISE	0.1 (5.38e-03)	0.113 (6.12e-03)	2.12 (6.23e-01)	0.826 (1.76e-01)
			$\bar{R}$	7.68	2.38	-	144.645
			$\bar{r}_1$, $\bar{r}_2$	5.53, 5.56	_	_	_
		0.4	AISE	0.102 (5.44e-03)	0.112 (5.64e-03)	4.18 (2.21e+00)	0.931 (1.76e-01)
			$\bar{R}$	7.34	2.22	-	146.855
			$\bar{r}_1$, $\bar{r}_2$	5.49, 5.49	_	_	_
	20	0.1	AISE	0.0637 (2.95e-03)	0.0706 (3.21e-03)	0.472 (8.01e-02)	0.304 (2.80e-02)
			$\bar{R}$	8.37	2.76	-	200.69
			$\bar{r}_1$, $\bar{r}_2$	5.81, 5.8	_	_	_
		0.4	AISE	0.0649 (3.06e-03)	0.0733 (3.30e-03)	0.484 (7.27e-02)	0.317 (2.53e-02)
			$\bar{R}$	8.24	2.78	-	206.16
			$\bar{r}_1$, $\bar{r}_2$	5.78, 5.78	_	_	_
200	10	0.1	AISE	0.0532 (1.98e-03)	0.0636 (3.12e-03)	2.33 (1.13e+00)	0.795 (2.98e-01)
			$\bar{R}$	8.48	3.02	-	191.175
			$\bar{r}_1$, $\bar{r}_2$	5.82, 5.82	_	_	_
		0.4	AISE	0.0548 (2.05e-03)	0.0686 (3.53e-03)	2.44 (1.17e+00)	0.828 (3.04e-01)
			$\bar{R}$	9.04	3.04	-	196.34
			$\bar{r}_1$, $\bar{r}_2$	5.71, 5.74	_	_	_
	20	0.1	AISE	0.0341 (1.43e-03)	0.0419 (2.02e-03)	0.301 (1.58e-02)	0.208 (4.50e-03)
			$\bar{R}$	8.99	3.74	-	318.645
			$\bar{r}_1$, $\bar{r}_2$	5.93, 5.92	_	_	_
		0.4	AISE	0.0348 (1.43e-03)	0.043 (2.22e-03)	0.328 (1.78e-02)	0.225 (4.74e-03)
			$\bar{R}$	8.01	3.6	-	327.395
			$\bar{r}_1$, $\bar{r}_2$	5.94, 5.93	_	_	_

Table S10: Simulation results for Setting 3 ($R = r_1 = r_2 = 4$) with the sparse design: see description in Table S8.

n	m	σ		mOpCov	OpCov	ll-smooth	ll-smooth+
100	10	0.1	AISE	0.105 (4.75e-03)	0.115 (7.58e-03)	24.1 (2.28e+01)	1.87 (1.19)
			$\bar{R}$	8.75	2.82	-	150.8
			$\bar{r}_1$, $\bar{r}_2$	5.26, 5.32	_	_	_
		0.4	AISE	0.11 (4.96e-03)	0.115 (8.33e-03)	26.2 (2.40e+01)	2.05
			$\bar{R}$	9.44	2.74	-	152.575
			$\bar{r}_1$, $\bar{r}_2$	5.37, 5.4	_	_	_
	20	0.1	AISE	0.0698 (2.74e-03)	0.0813 (4.63e-03)	0.614 (2.28e-01)	0.350 (8.35e-02)
			$\bar{R}$	6.63	3.24	-	210.515
			$\bar{r}_1$, $\bar{r}_2$	5.08, 5.14	_	_	_
		0.4	AISE	0.0721 (2.89e-03)	0.0859 (5.03e-03)	0.573 (1.74e-01)	0.344 (6.37e-02)
			$\bar{R}$	6.74	3.38	-	214.455
			$\bar{r}_1$, $\bar{r}_2$	5.11, 5.21	_	_	_
200	10	0.1	AISE	0.058 (2.62e-03)	0.0692 (5.33e-03)	0.454 (7.28e-02)	0.286 (2.89e-02)
			$\bar{R}$	6.26	3.12	-	182.74
			$\bar{r}_1$, $\bar{r}_2$	5, 5.06	_	_	_
		0.4	AISE	0.0598 (2.68e-03)	0.0733 (6.14e-03)	0.531 (1.07e-01)	0.323 (4.23e-02)
			$\bar{R}$	6.48	3.2	-	185.82
			$\bar{r}_1$, $\bar{r}_2$	4.99, 5.07	_	_	_
	20	0.1	AISE	0.0422 (1.37e-03)	0.0535 (2.64e-03)	0.267 (5.04e-03)	0.196 (3.59e-03)
			$\bar{R}$	6.29	4.49	-	332.09
			$\bar{r}_1$, $\bar{r}_2$	5.62, 5.69	_	_	_
		0.4	AISE	0.0424 (1.30e-03)	0.0494 (2.42e-03)	0.292 (5.30e-03)	0.212 (3.72e-03)
			$\bar{R}$	5.68	3.36	-	338.725
			$\bar{r}_1$, $\bar{r}_2$	5.59, 5.66	_	_	_